

**Developing Straits Times’s own**

**Predictive Advertisements using Analytics**

**BC2407 Analytics II: Advanced Predictive Techniques**

**Semester 2 Academic Year 21/22**

Team 3 S04

KOH I-ZACH (U2010476G)

POH SING YONG SHERLYN (U2011571H)

ROSHAN THAPA (U2022199F)

SNG YI XUAN (U2021009B)

APR, 2022

**Executive Summary**

Overview

The following report has been prepared to demonstrate the value of machine learning models in pop up advertisements for Straits Times Digital. The objective is to increase the click through rate (number of times an advertisement was clicked/number of times an advertisement was viewed) so as to increase revenue for Straits Times through increased commission fees. Furthermore, a first party predictive model will allow Straits Times to no longer rely on third party providers to take a cut of the commission fee. This report will present our research, analysis, action plans and detailed descriptions of how we intend to increase the click through rate using various machine learning models.

Situational Analysis

Straits Times currently uses Google Adsense and many other ad-serving platforms for its pop up advertisements on its app and web page. Every time a Straits Times reader clicks on a Google Ad displayed on Straits Times, Google takes about 32%-49% of the commission fee, while the rest of the revenue goes to Straits Times.

Key Issues

Subscribing to Google Adsense and allowing Google Ads to be run on the Straits Times webpage and app is convenient for Straits times but this may not be the most profitable. By developing its own in-house ad display algorithm, Straits Times would be able to take the full commission fee paid by advertisers.

Furthermore, Google’s recent announcement of removal of third party cookies means that they no longer have access to users’ data which suggests that their predictive advertisement will not be as effective as before. Straits Times readers would be receiving less targeted ads or ads that they would not have any interest in. These readers would not click on these ads and this would reduce the commission revenue Straits Times will receive from the advertisements.

Proposed Action Plans

We propose that Straits Times adopt a predictive model and develop an advertising platform themselves. The predictive model serves the purpose of providing Straits Times a decision making tool to decide whom they should target their advertisements to while the visualisation dashboard aims to share interesting insights to these advertisers through the use of graphs and charts.

Limitations

In order for our model to be successful, there is a need for more data to build a comprehensive model. Our data is based on the ads of Fitbit, an American smartwatch company. Realistically however, there will be many other companies that would want to advertise on Straits Times. We would thus need more data to train our model in order to predict our users' response towards their ads.

# 

**Content Page**

[**1. Introduction**](#_ibj80xgwfwgk)4

[1.1 Business Problem](#_rg95jy1vtgpn) 4

[1.2 Situational Analysis](#_vgebg3rhsrkx) 4

[1.3 Our Solution](#_uuaxm36dtb4r) 5

[**2. Case Justification**](#_evmv845ef2j1)5

[2.1 Why Straits Times?](#_tvimwl3e729k) 5

[2.2 Value of Predictive Analytics Model](#_uqa3qle2klrc) 6

[2.3 Information about Fitbit, a potential advertiser](#_d4k6rrg7h3us) 7

[Background](#_wc550alaaqeb) 7

[**3. Data Cleaning**](#_vqkbzpoe2gd2)8

[3.1 Handling Categorical Variables](#_nrgyah4gn1wc) 8

[3.2 Handling Continuous Variables](#_k7hj4zfmndrm) 8

[3.3 Dealing with Missing Values](#_tn4pfnjnnnu8) 8

[**4. Data Exploration**](#_3iq6b8rm6pi1)8

[4.1 Continuous Variables vs Response](#_izunwc649wum) 8

[4.2 Categorical Variables vs Response](#_skjblycizipj) 10

[**5. Analytics Solution**](#_t1w7av9g6kw5)11

[5.1 Proposed Approach](#_b2c58ewt3uvm) 11

[5.2 Association Rule](#_bnc23c96sy0a) 11

[5.3 Logistic Regression](#_jumsqjcmi3pt) 12

[5.4 CART](#_kyzju2zw7xc) 14

[5.5 Random Forest](#_hv2jkrhw3hmi) 15

[5.6 Neural Network](#_x61eoqzcdu7a) 15

[5.7 XGBoost](#_ke6ry8h1j34w) 17

[5.8 Model Recommendation](#_ki2iwelcaxkg) 18

[**6. Implementation Strategy**](#_y1964qhjhstx)21

[6.1 Pool of advertisers](#_153gj4tbyh9) 21

[6.2 Straits Times Adsense](#_isidefgwwst4) 21

[**7. Measuring Success**](#_lxetn13g0z3)22

[7.1 Metrics of Success](#_r5qfp825lyz5) 23

[7.2 Future Improvements](#_t8yk28t9e32b) 23

[**8. Conclusion**](#_158mia7xup7u)23

[**9. Appendix**](#_sz8ysavkokjx)24

[**10. References**](#_q5360qj8k0h0)42

# 1. Introduction

## 1.1 Business Problem

With rising privacy concerns in online advertising worldwide, there have been stricter privacy regulations enforced, leading to Google announcing a ban of 3rd party cookies by the end of 2023. It has instead promised to introduce a private alternative that users and advertisers will be happy with. There is therefore a need for a new solution that can match users to the most appropriate advertisement without the use of cookies. Straits Times, which uses many ad-serving platforms that require the use of cookies to advertise to their users, may now have to rethink ways to better identify which ads are the most suited to which user in order to ensure the same amount of clicks, hence revenue, as before. Additionally, these ad-serving platforms usually take up a larger portion of the revenue earned from the clicks on the advertisements. As a result, there exists an opportunity for Straits Times to innovate new ways to earn higher revenue without relying on such ad-serving platforms.

## 1.2 Situational Analysis

Google’s announcement of removing 3rd-party cookies

Advertisers will soon not be able to use third-party cookies for tracking users on Chrome. Digital advertisers will need to find new ways to attribute conversions, frequency cap ad placements, and retarget site visitors. Without proper tracking capabilities, advertisers will not be willing to invest as much in digital advertising. With digital advertising inventory valued lower due to lack of tracking, publishers will have to find other ways of generating revenue. (Olsson, 2022). Refer to [Appendix A](#a4ee2928gv1) for more issues on Google AdSense.

How Marketers and Advertisers are Reacting to Google's Phase-Out

Ad tech companies are now working on solutions to maintain the same type of performance and tracking alternatives for digital advertisers. In a recent survey. GetApp, which provided HubSpot with exclusive data, discovered that (Bump, 2021):

* 41% of marketers believe their biggest challenge will be their inability to track the right data.
* 44% of marketers predict a need to increase their spending by 5% to 25% in order to reach the same goals as 2021.
* 23% of marketing experts plan on investing in email marketing software due to Google’s new policy.

With the expected increase in spending among marketers and the rising need to find alternative solutions to targeted advertisements, there is a greater need for a solution that is able to help advertisers match their advertisements to their most appropriate ideal customer.

Glance into Straits Times’s Ad-Space

A quick look into Straits Times’s webpage, we observe that it is filled with advertisements from various ad-serving platforms. ([Appendix A](#qhvs0c1vnfj0)) With the removal of third-party cookies, it will be more difficult for these platforms to recommend advertisements to individuals as advertisers will not have access to an individual's browser history. This results in advertisers giving less targeted advertisements that will inevitably lead to lower clicks and conversions and therefore less revenue for both the advertisers and Straits Times.

Amount Straits Times earns from advertisements

Google pays Straits Times per every click on their ads, but it takes a commission. Publishers typically get **68%, or 51%** when using AdSense for reach. Depending on the niche, the commission can go from $0.20 to $15, with an average of $3 per click for publishers.

According to a Financial Times interview with Outbrain’s managing director in Europe, the network shares “about half” of any revenue generated with partner sites. The cost per click paid is said to range from $0.15 to $0.30 while click rates are in the neighbourhood of 0.50% to 0.75% (Johnston, 2021). Using those (big) ranges, we can come up with some revenue estimates:

* $0.15 CPC x 0.50% CTR x **50% revenue share** = $0.37 per thousand pageviews to publishers
* $0.30 CPC x $0.75% CTR x **50% revenue share** = $1.12 per thousand pageviews to publishers

As we can observe, most ad-serving platforms take up around 50% of the revenue from the clicks on the advertisements. If Straits Times could implement an alternative solution to benefit both advertisers and themselves, both parties have the opportunity to earn greater revenue.

Higher Reliance on First-Party Solution

The future of digital marketing would require many advertisers to rely on first-party solutions. First-party data is information companies collect about their customers that come from their own channels or sources. These are channels such as email campaigns, customer surveys, or website tracking (Olsson, 2022b). For example, creating an account with a site like Facebook or Google (in our case Straits Times), gives the site permission to track and save information on your activity. Instead of saving this information in a cookie, it's stored by the company and associated with your account.

With this information, Straits Times would have access to some of the reader’s personal details and activities. This information is valuable to understanding the behaviours and interests of their readers and allows advertisers to show more targeted advertisements to specific readers.

## 1.3 Our Solution

Using Predictive Analytics and data from Straits Times readers, we can develop a predictive model using machine learning that is able to predict whether a user will click through a specific advertisement. In essence, allowing Straits Times to continue being a profitable provider of ad space to their advertisers, whilst maintaining the confidence that the ads will continue to reach the ideal target audience. To show the application of our implemented solution, we will focus on Fitbit as our target advertiser wanting to advertise on Straits Times platform.

# 2. Case Justification

## 2.1 Why Straits Times?

According to Singapore Press Holdings (SPH), the company that owns Straits Times, print revenue has fallen significantly in the past 5 years. Across media businesses, advertising revenue has fallen greatly over the past decade due to the increased competition for digital revenue. This trend of falling advertising revenue is expected to continue. SPH and SPH media is thus continuously investing in the digital transformation of their business. Such investments include building up new consumer-facing digital platforms and products to strengthen the ability to design and monetise media content.

Straits Times also has a classified ads section dedicated to advertisements on both its print and digital newspapers. These pages are dedicated spaces for companies to advertise their products or jobs in a specific section of the newspaper and organised by product or service type. This differs from display advertisements that may contain pictures, texts, and logos. Display advertisements appear on the same page as the general news content instead of on a separate dedicated advertising page.

Straits Times’s first party-data

Straits Times has access to exclusive customer login information. When applying for a Straits Times subscription, the user will be prompted for information such as name, contact information as well as credit card information. Based on the users' past read articles, Straits Times is able to infer their interests, likes, and dislikes to be able to suggest advertisements to them.

Subscription and Usage

Straits Times is also a reliable news provider and is the most read paper in Singapore. According to a research company GfK, the Straits Times is accessed by 47% of Singapore’s population at least once a week, 45% of them subscribe to the e-paper. More people are also moving towards digital news instead of the traditional print media such as newspapers with 76% of ST readers on digital news sites. ([Appendix A](#ukvw78zetsk2))

Further, ST is read by people across different demographics, regardless of age and background. As the most-read local daily in Singapore with a diverse readership, Straits Times is an ideal platform for advertisers to promote products and services. ([Appendix A](#mze2itxf5qgt))

## 2.2 Value of Predictive Analytics Model

Higher revenue from digital advertisements

Rapid technological advances and the Internet have severely disrupted the traditional business model of media and advertising companies that relied on print advertising revenue. This has led to a decrease in global print advertising revenue by 7% year-on-year. In SPH, there has been a decline in newspaper print ad revenue, which fell S$28.6 million or 17.6 percent year on year. Print advertising revenue will continue falling as media consumption progressively shifts online. For example, digital circulation copies have grown 17.8% year on year. As such, there should be a greater emphasis on increasing the revenue from digital ads. This can be achieved through adopting its own digital advertising predictive analytics model, to garner greater revenue as a hefty share of their advertising revenue will no longer be paid to GoogleAds.

Increased click-through rates through more accurate targeting of customers

Predictive analytics allows Straits Times to target customers based on their needs instead of previous purchases. Without predicting consumer behaviour, ST can only speculate about the product that the customer will be interested in based on things that they have previously purchased. With predictive analytics, Straits Times can more accurately gauge customer interests.

Predictive Analytics is valuable to ST as it is able to use a reader's interest in a certain topic and suggest a product or campaign to them. This prediction determines if a reader will click on a digital display advertisement and thus improve its click-through rate. This also ensures that the ad campaigns are more competitive and attracts other companies to advertise on ST due to its greater outreach and higher click-through rate compared to other sites. A study conducted by the Aberdeen group suggests that there is a marked improvement in clickthrough rate from 4.5% to 7.9% when using predictive analytics. ([Appendix A](#5hl6ov6n7nbf))

Benefits of predictive analytical models to clients

In the realm of digital marketing, there are some key issues faced by marketing companies. For instance, one key issue is about effectively targeting the right audience. Predictive analytics can be used by the Straits Times website to help solve this as it will help connect these companies’ advertisements to the subscribers who are most likely to be interested in the products sold by the marketing companies based on the subscribers’ characteristics and interests. This will help to drive relevant traffic into these companies’ websites, through the increase in click-through rates. This will eventually increase the profits from the impressions on the ads. Another key issue is understanding customer behaviour. Through the predictive analytics model on the Straits Times website, it can provide some insights to the marketing companies with regards to customer and audience segmentation, new customer acquisition etc. This will provide feedback to client marketing companies, helping them craft more effective marketing advertisements.

Competitive Advantage

Furthermore, in the borderless online space, our local news media must now compete not just with international news organisations, but also entertainment providers and user-generated content. Thus, there is a need for SPH to gain a competitive advantage to stand out within the saturated news industry, especially given the small domestic market. By adopting its own predictive analytics model to increase click-through rates of ads on Straits Times’s website, SPH can increase its appeal to advertisers, beating its competitors in the local scene.

## 2.3 Information about Fitbit, a potential advertiser

## Background

The report uses Fitbit as an example of the advertisers that can benefit from the selected model. Fitbit is an American consumer electronics and fitness company. It produces wireless-enabled wearable technology, physical monitors and activity trackers such as smartwatches, pedometers and monitors for heart rate, quality of sleep and stairs climbed as well as related software. We will be focusing on Fitbit ads for their watches such as Fitbit Versa 3, Fitbit Sense etc.

Interest in boosting ad click-through rates

*(Please refer to* [*Appendix B*](#dj90dnd5wji9) *for a detailed breakdown of the figures)*

In 2020, Fitbit generated $1.13 billion revenue, a 20 percent loss year-on-year, and suffered a net loss of $190 million. While the number of Fitbit users have been increasing gradually in recent years, the Fitbit device sales has seen a decline with 10.6 million units sold in 2020, a 31 percent decrease year-on-year. Therefore, Fitbit will be interested in boosting its marketing efforts, especially in increasing its click through via displaying ads to its ideal target customers.

Variables of Interest

Based on research already done, we identified a few significant factors that affect ad responsiveness beyond factors related to the advertisement design. According to Pavlou and Steward (2000), additional factors such as the type of online activity and the user’s social context are important to users’ response to online advertisements. Since Straits Times has authorised access to customer information through its login function, certain personal data of interest to Fitbit can be tracked for machine learning. Apart from useful customer characteristics such as gender, age, education etc, other relevant data such as device type used, time spent on Straits Times application, number of times Straits Times application was opened can be used as predictors for whether a user will click on the Fitbit smartwatch ad. As Fitbit is a smartwatch that tracks health and fitness, we included variables such as number of articles related to health, technology and sports the Straits Times readers read because the congruence of the advertising message with the online content that is displayed in the same webpage was found to constitute a significant factor affecting the effectiveness of online advertising. (Papadopoulos, 2009)

# 3. Data Cleaning

Our dataset was simulated from a Fitbit smartwatch advertisement that was run on Straits Times. As Straits Times has a login function for users, we are able to obtain more personal data of the reader such as marital status and number of articles read. We assumed however, that the data was not fully accurate and thus we performed data cleaning. For the section below, we will be highlighting only the key steps we took during the cleaning process. For a more detailed data cleaning process, please refer to the Jupyter Notebook “Data Cleaning” attached together with the report.

## 3.1 Handling Categorical Variables

Referring to [Figure 3.1](#e8r3h8o8gumh), while looking through the categorical variables in the dataset, we discovered that there were irregularities in the data. The column “MaritalStatus” had a few wrong data entries. We used the number of children these people had to justify if we would replace them with “married” or “single”.

## 3.2 Handling Continuous Variables

While performing data cleaning on the dataset, it was found that there were unrealistic data entries [(Figure 3.2)](#605v1zb5mxxr). For example, the column “Age” had a data entry of 129 years old. This is incorrect as nobody has ever lived that long. We thus replaced this value with 29 under the assumption that the ‘1’ was inputted in by accident.

## 3.3 Dealing with Missing Values

The column income had some NA values. We replaced these NA values with the median income based on that person’s education level.

# 4. Data Exploration

In this section, we aim to conduct some preliminary data exploration on our variables. Below are some of the insights from the variables that are more interesting

## 4.1 Continuous Variables vs Response

|  |  |
| --- | --- |
| **Diagram** | **Findings** |
|  | The mean age is lower for those who clicked on the ads than those who did not click on the ads. The lower quartile and upper quartiles are also higher for those who clicked on the ads than those who did not. The relationship between these 2 variables appears to be strong. |
|  | The mean number of ads viewed is higher for those who clicked on the ads than those who did not click on the ads. The lower quartile and upper quartiles are also higher for those who clicked on the ads than those who did not. The relationship between these 2 variables appears to be strong. |
|  | The mean number of fashion articles read is higher for those who clicked on the ads than those who did not. Furthermore, the upper quartile is also higher for those who clicked. The relationship appears to be moderate. |
|  | The mean number of sports articles read is higher for those who clicked on the ad compared to those who did not. Furthermore, the upper quartile is also higher for those who clicked. The relationship appears to be moderate. |
|  | The mean number of sports articles read is higher for those who clicked on the ad compared to those who did not. Furthermore, the upper quartile is also higher for those who clicked. The relationship appears to be moderate. |
|  | The mean number of health articles read is higher for those who clicked on the ad compared to those who did not. Furthermore, the upper quartile is also higher for those who clicked. The relationship appears to be moderate. |
|  | The mean number of times the Straits Times app was opened by the user is higher for those who clicked on the ad compared to those who did not. Furthermore, the upper quartile is also higher for those who clicked. The relationship appears to be strong. |
|  | The mean number of ads viewed is higher for those who clicked on the ads than those who did not click on the ads. The lower quartile and upper quartiles are also higher for those who clicked on the ads than those who did not. The relationship between these 2 variables appear to be strong. |

## 

## 4.2 Categorical Variables vs Response

|  |  |
| --- | --- |
| **Diagram** | **Findings** |
|  | The proportion of people who clicked on the ads appears to be highest for phones, followed by tablets, laptops and computers. The difference is quite significant between the different device types so the relationship appears to be strong. |
|  | The proportion of people who clicked on the ads appears to be highest for undergraduates, followed by PHD, Graduate, Master, 2nd Cycle and lastly Basic. The difference is quite significant between the different education levels so the relationship appears to be strong. |
|  | The proportion of people who clicked on the ads appears to be highest for single people, followed by widows, divorced and married. The relationship appears to be moderate. |
|  | The proportion of people who clicked on the ads appears to be highest for people with 0 children, followed by 2 children, 1 child and 3 children. The relationship appears to be strong. |

# 

# 5. Analytics Solution

## 5.1 Proposed Approach

In this section, we will highlight the approach we will take to build the most optimal analytic model that will be able to predict whether a Straits Times reader will click on an advertisement publicized by Fitbit. Assuming that Straits Times has the past data of readers in response to a particular Fitbit advertisement, using variables such as “number of health articles read’, “average time spent on Straits Times”, “Income” etc ([variables that have been proven to be related to ad-responsiveness](#gw8hdtg9kms1)), we will be able to create an optimal model that - if given a new dataset of individuals, which of them will be predicted to click on the advertisement. Essentially, empowering Straits Times with a decision-making tool to help them decide who they should target the Fitbit advertisement towards.

Below is the approach we took to build the optimal analytical model:

1. Perform “Association Rules” to obtain a preliminary understanding of how different sets of input variables are associated with responding to the advertisement
2. Build different machine learning models where for each model:

* We will train the model on both an unbalanced and balanced dataset
* Plot the confusion matrix and calculate the different error rates

1. Compare results across all models and evaluate which is the most accurate and effective in the given context

## 5.2 Association Rule

Data Preparation

As only categorical variables can be used to create association rules, the continuous variables were converted to categorical variables and stored in new columns. A quartile was used to divide the data points of the continuous variables into 4 groups of similar size labeled 1,2,3,4, with 1 representing the bottom 25% and 4 representing the top 25%. Subsequently, the data was converted to transaction data format from a wide data format.

Associations Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Rules | Support | Confidence | Lift |
| {Education=Undergraduate,DeviceType=Tablet, Incomelvl=1,Timespentlvl=4} => {Response=1} | 0.01131 | 0.91176 | 4.1583 |
| {DeviceType=Phone,OpenApplvl=4,Timespentlvl=4} => {Response=1} | 0.024444 | 0.85897 | 3.9176 |
| {ChildrenHome=0,DeviceType=Phone,Timespentlvl=4} => {Response=1} | 0.029186 | 0.80000 | 3.6486 |
| {DeviceType=Phone,Techlvl=4,Timespentlvl=4} => {Response=1} | 0.017147 | 0.85455 | 3.8974 |
| {ChildrenHome=0,DeviceType=Phone,OpenApplvl=4,Timespentlvl=4} => {Response=1} | 0.013864 | 0.90476 | 4.1264 |

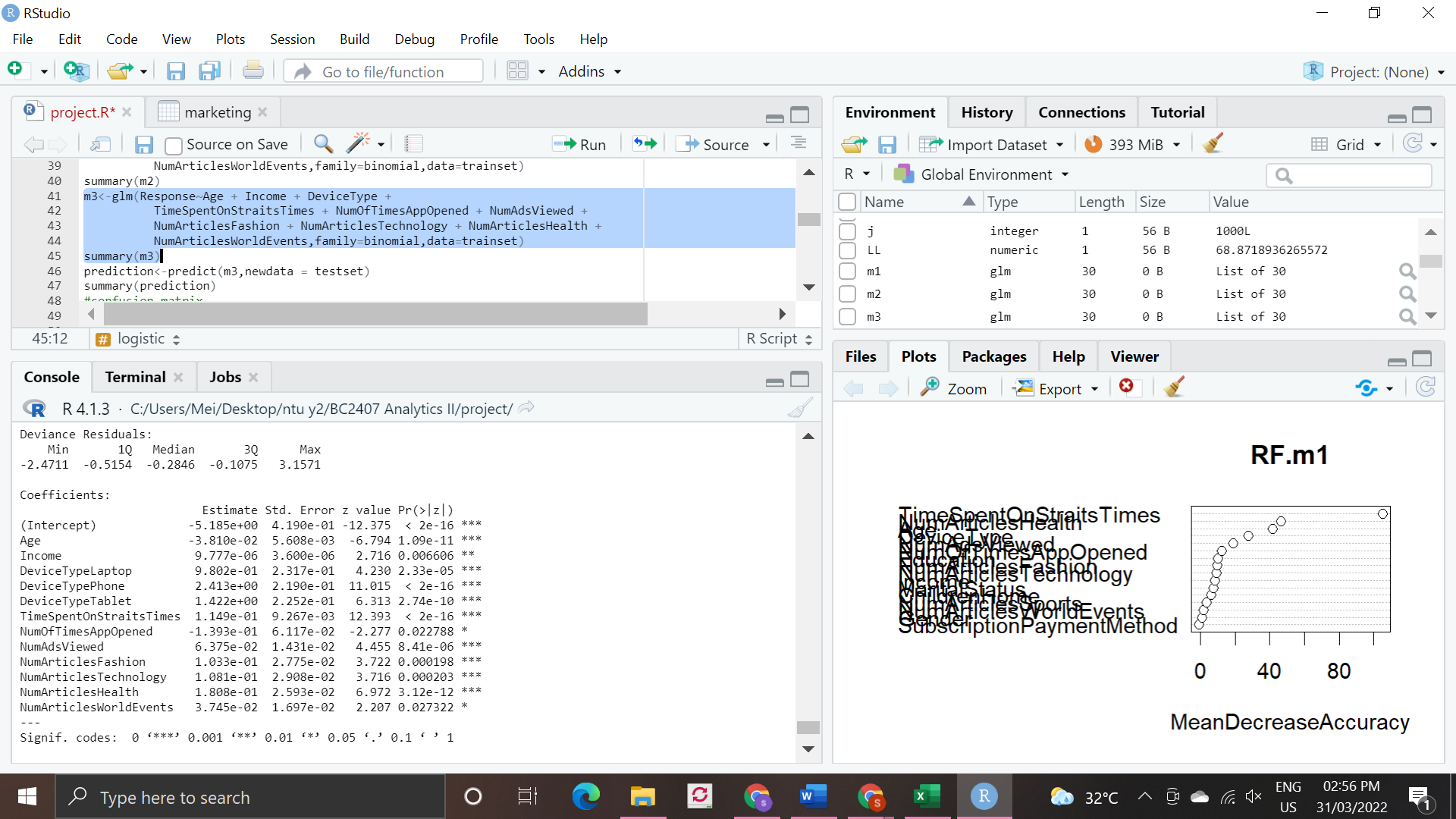
Support only shows how often you will be able to apply the association rule, whereas confidence only tells you the strength of the rule in situations where it is applicable. Thus, they complement each other and should be used together. As such, we decided to use lift to rank the association rules as lift measures how useful the rule is in the context of the situation. We chose 4 interesting association rules with one of the highest lift values and relatively high confidence and support values among those with the highest lift values. Generally, a high lift value above 1 means that the rule helps to boost the probability of a user clicking on the ad.

From these association rules, we can identify the important relationships between variables and users clicking on the ad and pinpoint some common strong predictors of a user’s response, which are Time spent on Straits Times Ad being high, Number of Times the Straits Time App was opened being high and Device Type being phone. These give valuable insights into the characteristics of users whom Straits Times should target to display the Fitbit ads to as they are the ones who will most likely click on the ad. In general, since the variables related to users’ online activity were the strongest predictors, Straits Times can reuse these variables as predictors for future training models on new clients.

## 5.3 Logistic Regression

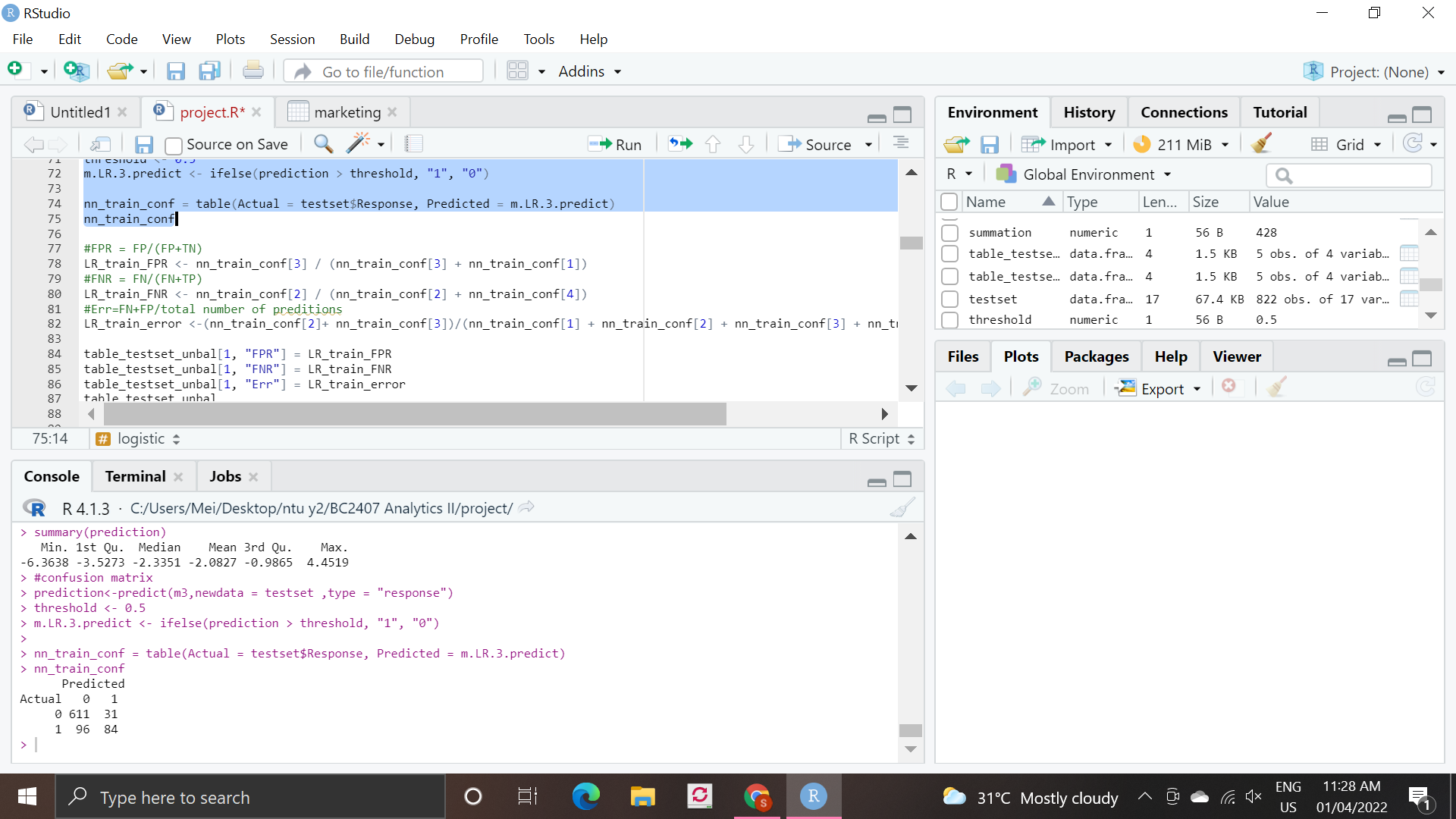
Logistic regression is a process of modelling the probability of an event occurring and is used to establish the relationship between a dependent variable and several independent variables. Backwards elimination is a variable selection method that begins with all variables which are subsequently removed based on lowest statistical significance.

Training on Unbalanced Dataset



*Figure 5.3.1: Results obtained from logistic regression with BE on the unbalanced trainset.*

Figure 5.3.1 shows the variables deemed to have the highest statistical significance. The variables were obtained through conducting logistic regression with backwards elimination on the unbalanced dataset. The statistically insignificant values were removed until the model contained only statistically significant variables. With this method, we determined that the statistically significant variables are: age, income, the type of device used to access Straits Times online, time spent on Straits Times, number of times the Straits times app was opened, number of ads viewed, number of articles read about fashion, technology, health and world events. Out of these, income, number of times the app was opened and the number of articles read about world events have lower statistical significance compared to the other variables. In general, most variables have a positive correlation with response with the exception of age and number of times the app was opened.



*Figure 5.3.2: Confusion matrix from logistic regression analysis on the unbalanced trainset.*

With the resultant model, we were also able to plot the confusion matrix as shown in Figure 5.3.2. From this, we determine that the model has a false positive rate (FPR) of 0.048, false negative rate (FNR) of 0.533 and classification error (Err) of 0.154. The model is deemed to have a low FPR, and low classification error. The significant difference between the values of FPR and FNR, the model suggest that the model has a greater tendency towards one response. As such, for a more nuanced result, we performed the analysis on a balanced dataset.

Training on Balanced Dataset

The same regression analysis with backward elimination was performed on the balanced dataset. It has been determined that the statistically significant variables are: age, income, the type of device used to access Straits Times online, time spent on Straits Times, number of times the Straits times app was opened, number of ads viewed, number of articles read about fashion, technology, and health as shown in [Figure 5.3.3](#uf88nliedzwr) Out of these, income, number of times the app was opened and the number of articles read about health and technology have lower statistical significance compared to the other variables.

From the confusion matrix in [Figure 5.3.4](#6enaa0q91jwp), we then determined that the model has FPR of 0.176, FNR of 0.205 and classification error of 0.182. Using the balanced trainset, the classification error has increased marginally, while FNR has decreased. FPR has also increased by 0.128. This is due to FPR and FNR having a similar error rate when trained on a balanced dataset.

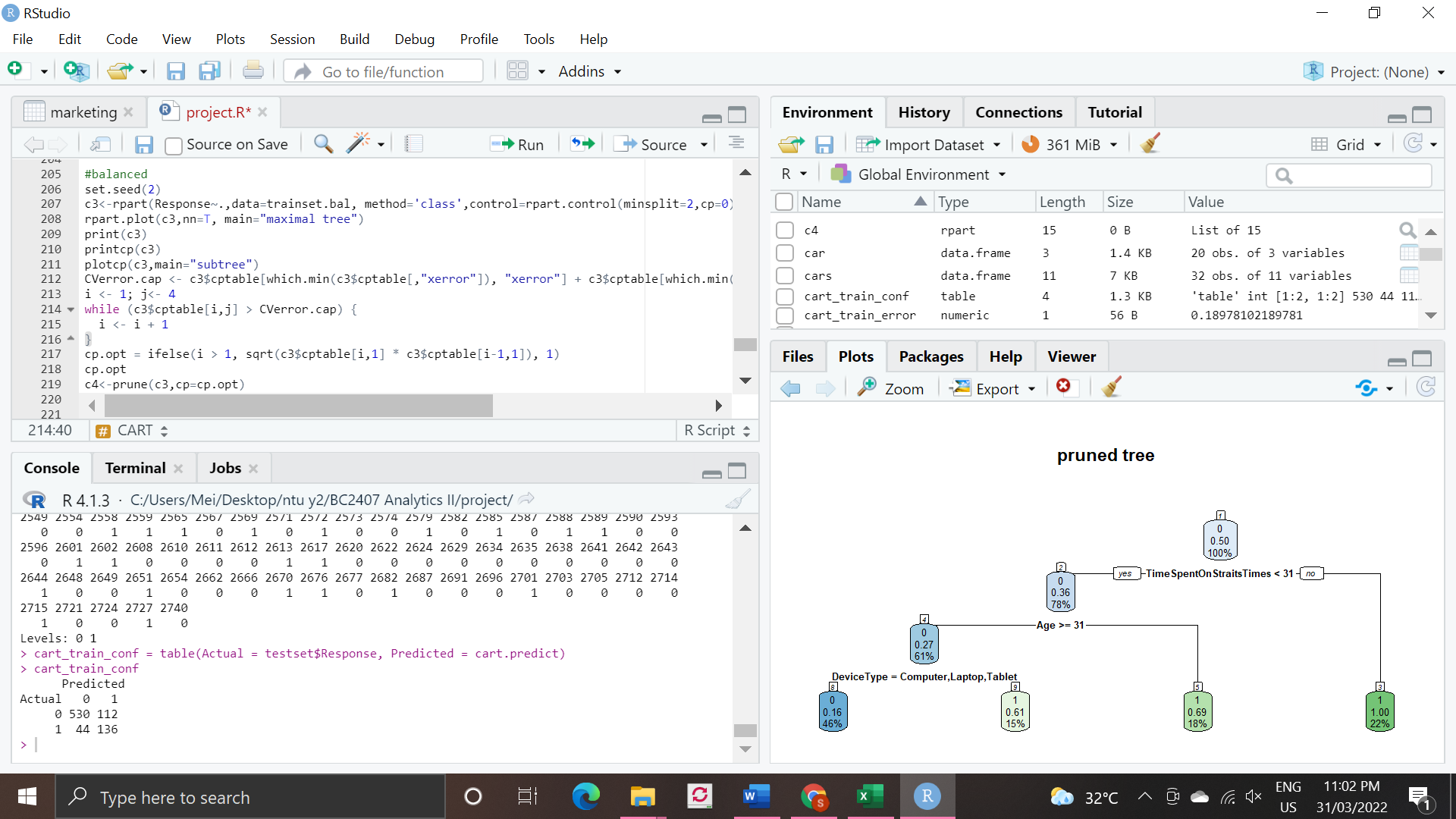
## 5.4 CART

CART is short for Classification and Regression Tree which shows how the value of a target variable is predicted based on the value of other variables.

Training on Unbalanced Dataset

Conducting CART analysis on the unbalanced trainset produced the decision tree shown in [Figure 5.4.1](#s7r21y8f3ipe). We determine that the optimal tree has 2 levels based on the predictors time spent on Straits Times and the number of articles related to health read. This means that the model will predict that the reader clicks on the ad if he spends less than 31 minutes on Straits Times and has read more than 9 articles related to health. 3% of total number of readers who exhibit these characteristics will be predicted to click on the ad. 9% of the total spend more than 31 minutes on Straits Times, and still predicted to click the ad. From the confusion matrix in [Figure 5.4.2](#vbrwsw7g4000), we can determine that the model has FPR of 0.009, FNR of 0.550 and classification error of 0.128.

Training on Balanced Dataset



*Figure 5.4.3: Decision tree from training model on balanced trainset*

From the decision tree shown in Figure 5.4.3, we determine that the optimal tree has 3 levels based on the predictors time spent on Straits Times, age and the type of device that the reader accesses Straits Times from.This means that the model will predict that the reader clicks on the ad if he spends less than 31 minutes on Straits Times, whose age is greater than or equal to 31 and does not use either a computer, laptop or tablet to access straits times. 15% of the total number of people who display the above characteristics are predicted to respond favourably to the ad. 18% of the total who are younger than 31 years old and spend less than 31 minutes on straits times are predicted to click the ad. 22% of the total are people who spend more than 31 minutes in Straits times and also predicted to click on the ad.

These variables differ from training the model on the unbalanced trainset. However, the prediction is deemed to be balanced when applied on a balanced trainset as there is an equal opportunity for the model to be trained on both predictions (response = 1 and response = 0).

From the confusion matrix in [Figure 5.4.4](#dchn537r02rr), we determine that the model has FPR of 0.238, FNR of 0.127 and classification error of 0.214. As with logistic regression, when using the balanced trainset, the classification error has increased marginally, while FNR has decreased. FPR has also increased by 0.086.

## 5.5 Random Forest

Random Forest (RF) is an algorithm that makes predictions by creating and combining multiple decision trees.

Training on Unbalanced Dataset

From the random forest model we can plot the confusion matrix shown in [Figure 5.5.1](#kix.ycsn4vr69zm4). From this, we can determine that the model has FPR of 0.031, FNR of 0.461 and classification error of 0.125. As before, we will conduct the analysis again on the balanced trainset for a nuanced result.

Training on Balanced Dataset

From the new random forest model we can plot the confusion matrix shown in [Figure 5.5.2](#kix.7hu6rjmxe27p).. From this, we can determine that the model has FPR of 0.151, FNR of 0.122 and classification error of 0.145. We observe that the FPR increases by 0.12, classification error increases slightly and FNR falls. This observation is consistent with the other models explained above. This observation is also aligned with that of typical balanced datasets where we expect a slight increase in error and FPR, with a significant decrease in FNR. At this stage, the random forest model produces the most optimal result.

Variable importance

Conducting the analysis using the balanced dataset produced the plot in [Figure 5.5.4](#x5jo2ra9qpyu). We observe that the top 8 most important variables are time spent on Straits Times, Age, device type, number of articles on health read, number of ads viewed, number of children in the household, marital status and income. This is a marked difference from the unbalanced trainset with 3 of the 8 variables differing from the model run with the unbalanced trainset and with different levels of importance. These variables will later be used as inputs in neural network analysis.

## 5.6 Neural Network

Neural Network (NN) is a type of deep-learning model that comprises node layers, an input layer, one or more hidden layers, and an output layer. Each node connects to another and has an associated weight and threshold. The result from the output layer would be a prediction that can be used for both continuous and categorical variables

Data Preparation

We will be using R’s “neural net” library for our NN model, which allows us to plot the neural diagram. However, a limitation of this library is that it is not able to handle categorical variables. As such, before we train our NN model, we first converted all our categorical variables into dummy variables [(Figure 5.6.1)](#jgwib62kn2d6)

Training on Unbalanced Dataset

1. Default Settings (Without Normalization)

We first performed an iteration with 1 hidden layer and 2 nodes in the network, trained on the unbalanced dataset. We observed that the confusion matrix obtained [(Figure 5.6.2)](#k7orn3m0bmr5) is not ideal as the model fails to predict any responses (Response = 1). It has an FNR of 1 and a classification error of 0.219. This suggests that the Neural Network model that is not normalized may be unusable as it is not able to predict any individuals that will respond hence providing no value to Straits Times. Additionally, if we were to plot the generalized weights plot for some of the variables [(Figure 5.6.3)](#5ntm8p6bqmg5), we can also observe that for most of the variables, the GW shows a constant horizontal 0 which suggests that the variable is considered to have no effect on response. Therefore, much optimization is required henceforth.

1. Normalization with different numbers of neurons (Unbalanced dataset)

To optimize our model, we first normalize our continuous variables by performing a MIN-MAX normalization. Next, we proceeded to run several iterations of the model, each time varying the number of nodes in the 1st hidden layer. From our research, if our data is less complex and is having fewer dimensions or features then a NN with 1 to 2 hidden layers would work (Sachdev, 2021). Additionally, the number of nodes in each layer should be between the size of the input layer and the output layer. The most appropriate number of hidden neurons is: sqrt(input layer nodes \* output layer nodes) = **sqrt(16 \* 1) = 4.** Therefore, to further optimise the NN model, we will tune the number of hidden nodes in a layer and find the one that results in the lowest error.



*Fig 5.6.4: Table showing the error rates for different iterations of the NN model (Unbalanced)*

From the above table, we observe that indeed with normalization, the error rate has decreased and FNR is no longer 1, thus the model is now able to predict Responses = 1, showing a drastic improvement from the model without normalization. Additionally, 4 nodes in a layer might be the most optimal as shown from the model having one of the lowest Err, FNR, and FPR (which concurs with the finding above). Here note that we did not choose the model with the lowest Err as the best model. Instead, we chose the model that is the most balanced as the most optimal one. More will be explained in the later section

Training on Balanced Dataset

Next we proceeded to train the model on the balanced dataset, running an iteration with just the balanced dataset while another with the top 8 variables from RF as input variables trained on the balanced dataset. After which, we will compare their effectiveness.

1. Comparison between NN trained on balanced + Inclusion of VarImpt from RF

## 

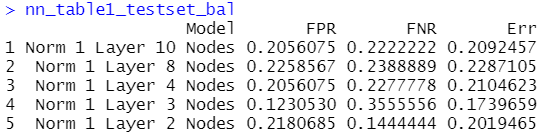
*Figure 5.6.5: Table summarising the error rates between NN balanced vs NN balanced + VarImpt*

Referring to the above table, we can make several key observations:

1. Training the model on the balanced dataset results in a general increase in Err. FPR and FNR are now more balanced
2. NN trained on the balanced dataset with just the top 8 variables from RF, results in a more accurate and balanced model than just the model trained on the balanced dataset
3. With normalization, GW plots show that the weights are no longer 0 suggesting that the variables now do have an effect on response [(Figure 5.6.6)](#kix.yvuraqm29ya3)

For a visual representation of the neural network diagram plot for the model that is trained on the balanced dataset with input variables from RF, please refer to [Figure 5.6.7](#kix.naeskqa9bp5v)

1. Normalization with different numbers of neurons (Balanced + Input Variables from RF)

****

*Figure 5.6.8: Table showing the error rates for different iterations of the NN model (Balanced with Input Variables from RF)*

From the above table, we observe that the NN model with 2 nodes in the layer might be the better model for NN trained on balanced with input variables from RF. This is because it has the 2nd lowest Err as well as a generally low and balanced FPR and FNR.

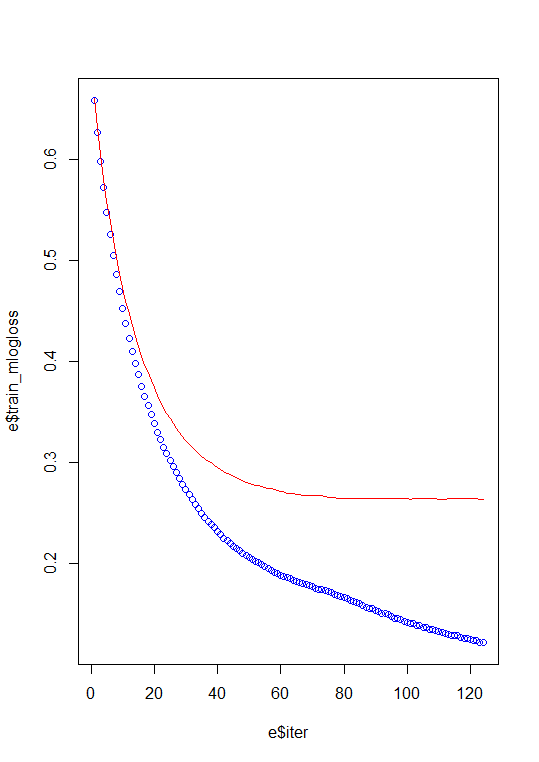
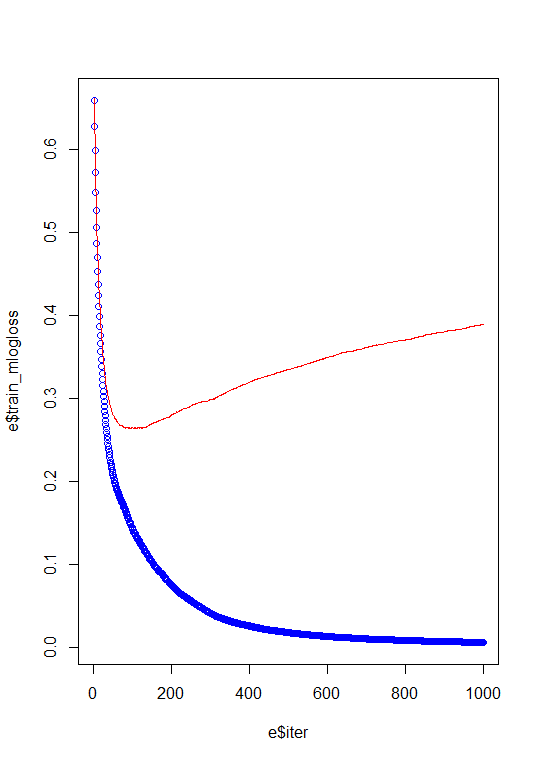
## 5.7 XGBoost

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework that can be used to solve regression and classification problems. Similar to all the above models, we will first train the model on the unbalanced dataset followed by on the balanced dataset, and then compare their results.

Data Preparation

For XGBoost, similar to NN, it is unable to handle categorical variables. Therefore, we first converted all the categorical variables into dummy variables using a matrix [(Figure 5.7.1)](#kix.uwj4zfmewhq6)

Training on Unbalanced Dataset



*Figure 5.7.2: Train and test log loss graph comparison*

We first performed XGBoost with the default parameters, setting the total number of iterations to 1000 initially and the learning rate (eta) to be 0.05. From the graph above, we observe that as the number of iteration increases, the train loss always decreases (blue line). On the other hand, the test loss (red line) first decreases and then increases after a certain point, indicating that the model is overfitting from that point on. Hence, we find the point at which the test loss is the minimum and repeated the algorithm, by this time, setting nrounds = 124 (the 124th iteration results in the point with the lowest test error)

From the confusion matrix [(Figure 5.7.3)](#kix.v1ive5s5p9qo), we observe that the Err rate is 0.123, FPR is 0.037 and FNR is 0.428 for the model trained on the unbalanced dataset. The result obtained shows that XGBoost has currently the lowest Err rate, while RF is a very close 2nd. This suggests that XGBoost currently, has the highest predictive accuracy. However, as mentioned, there is still quite a distinction between FPR and FNR thus, we will train the model again but this time on the balanced dataset

Training on Balanced Dataset

Repeating similar steps as explained earlier. We first ran the model with 1000 iterations, found the ith iteration where the test loss was the minimum, re-run the model again with that ith iteration, and plotted the confusion matrix. Referring to the confusion matrix [(Figure 5.7.4)](#kix.292scxwf70tc), our classification error has now increased slightly to 0.152, FPR has also increased to 0.139 but FNR has decreased greatly to 0.200. Similar observations are made, the model that is trained on the balanced dataset will generally show a very slight decrease in general predictive accuracy, however, the model is now better able to predict both classifications of response.

Variable Importance

Referring to [(Figure 5.7.5)](#kix.tbz7ici7ozli), we observe that the top variables from XGBoost, in decreasing order, are “TimeSpentOnStraitsTime”, “Age”, “NumArticlesHealth”, “NumsAdsViewed”, “NumOfTimesAppOpened”, “NumAdsTech”, “NumAdsFashion”. Comparing these variables with the variables that are important from Random Forest, we see that many of the variables are repeated suggesting that these variables are consistent and are generally the variables that Fitbit and Straits Times should focus on

## 5.8 Model Recommendation

Before we evaluate which of these 5 machine learning models would be the best, we first have to understand what our error rates signify in the context of Straits Times.

Understanding FPR, FNR, and Err

False Positives refer to users who were predicted to click on the ad but actually did not click on the ad. A large number of false positives has many business implications for Straits Times. Firstly, a large number of false positives indicates the inherent weakness of the predictive model. The objective of our model is to present ads to a Straits Times user that he/she will click on. If the reader does not click on the ad, our model objective has not been fulfilled. Secondly, a large number of false positives indicate that there has been revenue that has been missed out. Ads that the reader would have clicked on could have been shown instead. Thus, the potential revenue loss is represented by this opportunity cost.

False Negatives refer to users who were predicted to not click on the ad but in actuality, they will click on the ad. Since our business solution to Straits Times is to provide them with a decision-making tool to help them to decide who they should show Fitbit’s advertisement to. Thus, if our machine learning model predicts a reader to not click on the ad, Straits Times will not display the ad to the individual. Therefore, should our model have a large number of false negatives, this suggests that we are missing out on a large portion of revenue. This is because, for all the readers that we do not show the advertisement to, if they had been instead shown this advertisement, they would have actually clicked on it, thus Straits Times is losing revenue by not showing the advertisement to these individuals.

Lastly, as per all predictions, a low misclassification error is definitely more favorable. A low classification error suggests that the prediction is able to accurately predict the outcomes with high certainty. Classification error describes the overall predictive power of the model and hence it is the error rate that underpins all other error rates.

Determining the relevant metrics

When comparing FPR and FNR, we should note that the size of the groups predicted to click versus the size of the group predicted to not click, will affect the revenue of Straits Times. For the following discussion, we will classify readers that are predicted to click as Group 1 and readers to not click to be as Group 0. Suppose Group 1 is larger in size (100 readers) as compared to Group 0 (10 individuals). And suppose that both FPR and FNR is 10% and the revenue Straits Times would have earn from a successful click is $1.

This means that Straits Times would have been able to earn $90 ((1-FPR) \* 100 readers) from the readers that were predicted to click. On the other hand for Group 0, Straits Times would only be losing out on $1 (FNR \* 10 readers) of potential revenue. If now both FNR and FPR increases by 10%, this means that Straits Times would now only earn $80 ($10 decrease in revenue) from Group 1 and would be losing out on only $2 of potential revenue from Group 0 ($1 decrease in potential revenue). Therefore, we can observe that depending on the sizes of the groups, FPR and FNR would stamp a different weightage on revenue. If Group 0 is greater in size, a high FNR would result in a greater loss of potential revenue while if Group 1 is greater in size, a high FPR would result in a smaller revenue.

In addition to just FPR and FNR. We could also take into account the model’s precision and recall. Precision is the ratio between the true positives and all the predicted positives (TP / TP + FP) while recall is the ratio between true positive and actual positives (TP / TP + FN). A high precision implies that the predictive model is accurate in selecting the readers that will click on the ad while a high recall suggests that out of all the actual readers that had clicked on the ad, the model had correctly identified a large proportion of these readers. In our case, Straits Times would want both a high precision (higher predictive accuracy) and high recall (better market capitalization). In cases where both metrics are important, we can use a metric F1-score (Refer to [Figure 5.8.1](#lkju77jvp6y6) for F1-score formula)

Hence, for Straits Times, it is optimal that our model should have a low FPR and FNR, thus a low misclassification error. Additionally, it should have a good balance between precision and recall and hence a high F1-score.

Comparison of metrics between ML models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | | **FPR** | **FNR** | **Err** | **F1-Score** |
| Unbalanced | Log Reg (BE) | 0.048 | 0.533 | 0.155 | 0.569 |
| CART | 0.009 | 0.550 | 0.128 | 0.607 |
| RF | 0.031 | 0.461 | 0.125 | 0.653 |
| NN | 0.069 | 0.417 | 0.145 | 0.638 |
| XGBoost | 0.037 | 0.428 | 0.123 | 0.671 |
| Balanced | Log Reg (BE) | 0.176 | 0.206 | 0.183 | 0.656 |
| CART | 0.238 | 0.128 | 0.214 | 0.641 |
| RF | 0.151 | 0.122 | 0.145 | 0.726 |
| NN | 0.218 | 0.144 | 0.202 | 0.650 |
| XGBoost | 0.139 | 0.200 | 0.152 | 0.697 |

Finally, we make a comparison between the 5 different machine learning models for both the unbalanced and balanced dataset.

1. Unbalanced Dataset

We observe that Random Forest and XGBoost have the lowest Classification Error and highest F1-score. Both also generally have one of the lowest FPR and FNR. With an Err of about 0.12, both models generally have a strong predictive accuracy. However, we can clearly observe that for all models trained on the unbalanced dataset, there is quite a disparity between FPR and FNR with FNR generally being way higher than FPR.

1. Balanced Dataset

As the FPR and FNR are quite different when trained on the unbalanced dataset, we trained the models now on the balanced dataset so that the models have equal opportunity to learn from both classifications of responses. Below are some of our key observations:

1. Training the model on the balanced dataset generally leads to a very slight increase in Err, suggesting a very slight decrease in predictive accuracy though not significant.
2. F1-Score generally increases for all models trained on balanced dataset which suggests a high accuracy for the dataset
3. FNR will decrease significantly while FPR increases slightly. However, both FNR and FPR will generally be more balanced, indicating that the model is better able to predict responses of both classifications
4. Random Forest has the best results among all the models. It has the lowest Err rate of 0.145, lowest FNR of 0.122, and the 2nd lowest FPR of 0.151, a very close 2nd only to XGBoost. Additionally, it also has the highest F1-Score, indicating it is also the most accurate in terms of accuracy for the current dataset

From our results, it is quite clear that Random Forest (trained on the balanced dataset) is the best model in terms of Err rate, F1-Score as well as having a balanced FPR and FNR. XGBoost is a very close 2nd. Logistic Regression with BE is next, followed by Neural Network, and lastly CART.

Model Considerations

Before we conclude on the final model to recommend to Straits Times, we should also consider other factors aside from metrics such as the limitations of the models, which could affect the final choice made by Straits Times.

Aside from the error rates, Neural Network is not a recommended model for Straits Times because NN requires a large amount of time to optimize and run. There exist many hyperparameters in the model that can be tuned to better improve the accuracy of the model hence, in the context of Straits Times, where predictions are required on the fly and in large amounts, NN is definitely not ideal. Additionally, NN is a model with low explainability, it is difficult to explain why and how the model works, and hence it does not help with the explanations of the variables and responses which are crucial knowledge to have for both Straits Times as well as potential advertisers on the platform.

Our chosen model thus ought to have not just a low misclassification error and high F1-Score, but it should be relevant and useful in the context of Straits Times as well. We hence recommend developing Strait’s Time predictive advertisements using Random Forest, ideally trained on a balanced dataset. Random Forest is ideal because it is robust to outliers and runs efficiently on large datasets without need for much data cleaning. It has a low risk of overfitting and has one of the best accuracies when it comes to classification problems. Thus, in the context of Straits Times where they exist many readers at a time, and predictions are required in large amounts, Random Forest will be able to handle such a task.

# 6. Implementation Strategy

Overall, we want a seamless experience for companies to advertise with Straits Times. Besides the use of our predictive model to increase click through rate for these advertisers, we also want to develop a visualisation dashboard for them so that we can share the insights of our models and help the advertisers make better advertising decisions.

## 6.1 Pool of advertisers

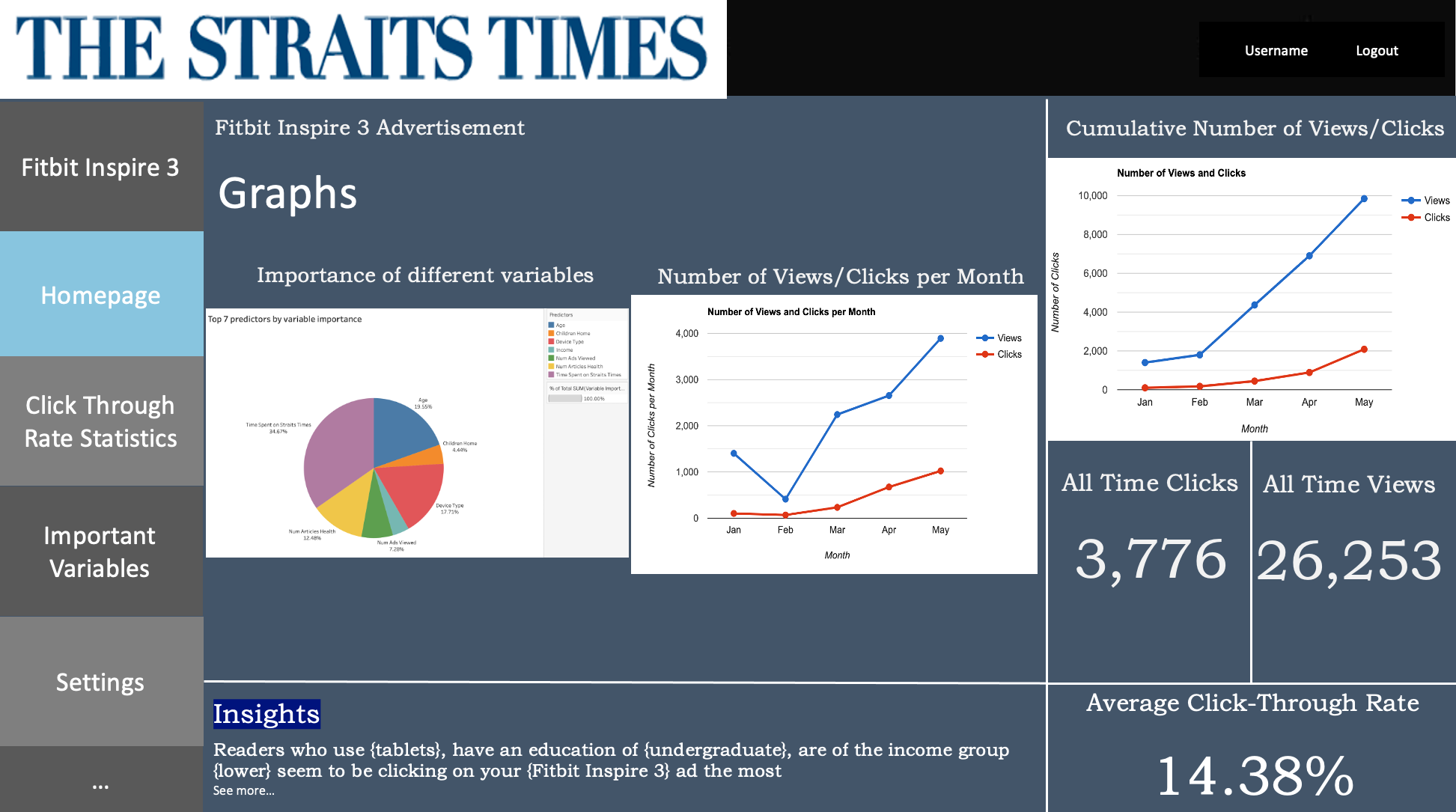
Straits Times currently has a large pool of advertisers who display ads on its hardcopy newspapers as well as its e-papers (PDF file of the hardcopy version). These advertisers could be offered a trial period to use Straits Times Adsense to have pop-up ads in addition to their ads on the hardcopy version/e-paper. During this trial period, we hope to demonstrate the effectiveness of our predictive model in increasing click through rate for their pop-up ads.

## 6.2 Straits Times Adsense

Straits Times Adsense is a multi-solution/all-in-one platform which allows advertisers to access everything related to their advertising campaign on our platform. Advertisers will use this platform to decide how much money they want to put into advertising their campaign on Straits Times as well as how many people have both seen and clicked on their ad. Besides this, the platform would also serve as a visualisation dashboard for these advertisers. They will be able to obtain insights on the model through various graphs and charts on the progress of their advertising campaign. Advertisers can use these insights to reformulate their marketing campaign or business strategy.

Visualization Dashboard

*(For a more detailed picture of the piechart, please refer to* [*Figure 6.1*](#mb5r7oz6bui)*)*



*Figure 6.2: Homepage of Adsense dashboard*

The figure above shows how the homepage of our adsense dashboard will look like. Real-time click through rate, number of people who have seen the ad, number of people who have clicked the ad and the click through rate of a particular Fitbit ad will be shown on the homepage as well as a chart of these statistics over the months. In addition to this, we want to visualise interesting insights to the advertiser. In the case of Fitbit above, the dashboard would show the top 7 variables that are important, time spent on Straits Times App, age, type of device, number of articles on health read, number of ads viewed, children at home, number of times Straits Times app was opened and income.

Referring to [Figure 6.3](#32xa3tmrgll0) and [Figure 6.4](#sfj6x1oqdcfe), the analytical dashboards act as visual aids to give a general overview of how the 7 most important predictors interact with each other. They present data in a clear way that facilitates decision-making for the users who consist of both Fitbit and Straits Times. Historical data was analysed to visualise trends and discover insights regarding the 7 most important predictors.

Utility of the dashboard

The purpose of the dashboards is to help Fitbit to better understand Fitbit’s ideal customer base on the Straits Times website. By analysing the characteristics and behaviours of people who clicked on the ads, Straits Times can better advise Fitbit on how to craft effective and more informed marketing strategies targeted at people who are more likely to click on the ad. Through this, Fitbit can improve their advertisement designs, leading to an increase in click-through rate, resulting in more revenue from advertisements for both Fitbit and Straits Times.

Additionally, by better understanding the target customer characteristics of each client company, Straits Times can be more effective in their display of advertisements to their users, ensuring higher click through rates on the ads in their websites, gaining higher revenue. Straits Times can more efficiently utilise their resources, avoiding the display of ads to users who will most likely not click. This will help to decrease opportunity cost from unnecessary use of resources and increase overall profits. Over the long run, the continued effective click through rates through the use of machine learning models will attract more new advertising clients and retain current ones. Furthermore, Straits Times can also use these visualisations to predict responses for future new clients with profiles similar to Fitbit with no existing historical dataset.

# 7. Measuring Success

## 7.1 Metrics of Success

Click Through rate before and after predictive model is implemented

The main objective of the project is to increase the click through rate of ads. The current average click through rate for most online ads stands at about 0.50%-0.75%. After running our model with Straits Times, our model should be able to improve this rate. This increase in click through rate would be an important metric of our model’s value to Straits Times.

Number of advertisers that choose to advertise with us

Being a new model, we do not expect to have many advertisers that choose to run their ads on our platform at first. However, if our model can be shown in the market to work effectively in increasing click through rates, the popularity of our model will increase and this should lead to an increasing trend of advertisers using Straits Times Adsense. Thus, the number of advertisers that choose to use Straits Times Adsense would be an important indicator of the popularity of our predictive model in the market.

Profit earned before and after predictive model is implemented

Another key metric for success is the extra profit earned for Straits Times with this new model. Currently, Straits Times uses Google Adsense, which takes about half of the commission fee generated when a user clicks on an ad. After deploying our predictive model, we expect an increase in overall revenue as a cut of the commission fees generated would no longer be taken by Google. A significant increase in profit for Straits Times would be an important indicator of the importance of our model to Straits Times.

## 7.2 Future Improvements

Train models on new data

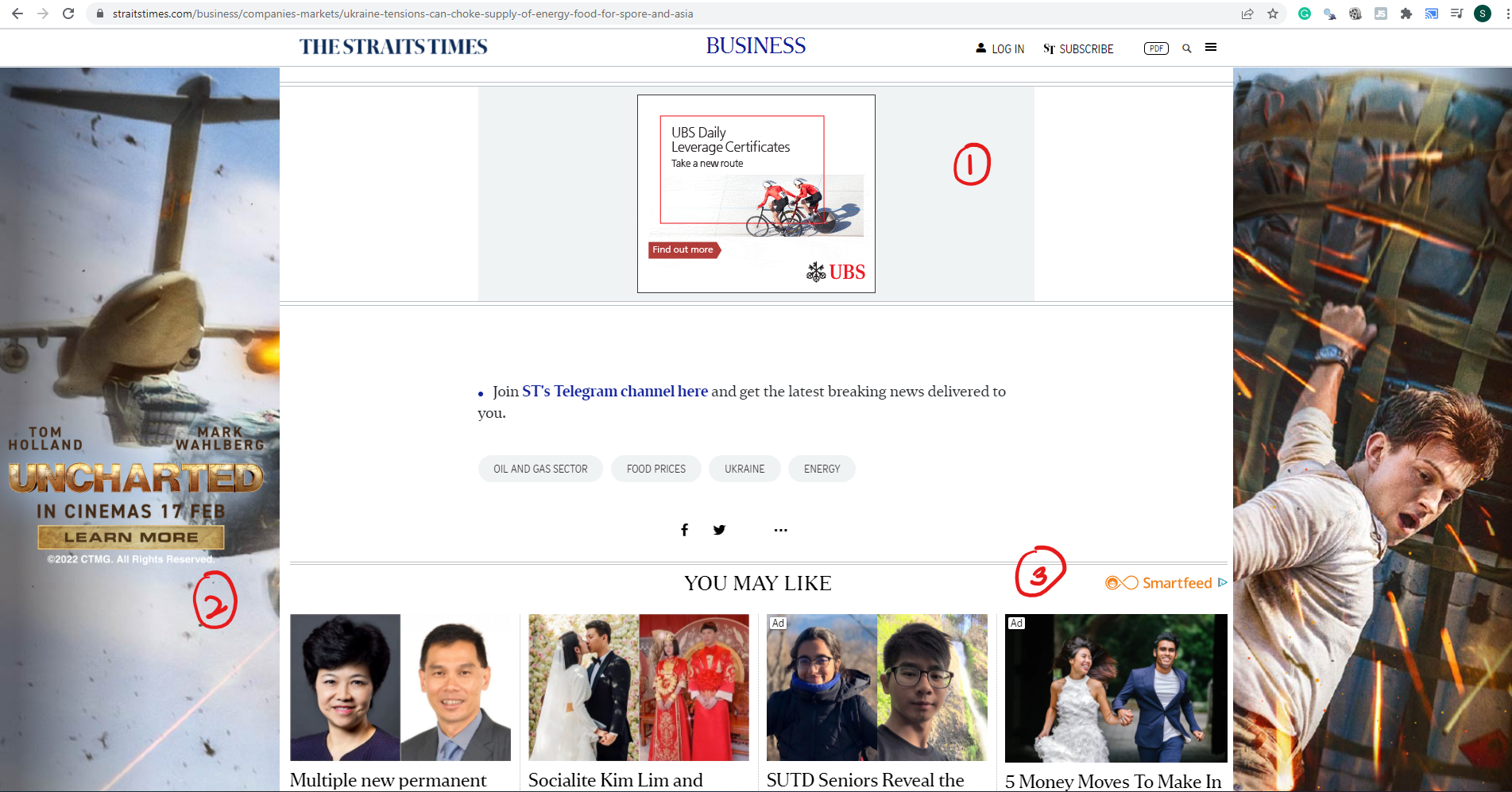
With the ever-changing demographics of users and the rise and fall of global trends, it is crucial that our predictive model is optimised to learn and to pick up these changes. Hence, it is important that our model is constantly being fed with new data so that the model can be better trained to predict the responses of future users. Additionally, currently, we have only run our model for the advertiser FitBit, this may not be sufficient in demonstrating the overall success of the model. Only by running it against a wide range of advertisers will we then be able to see the true accuracy and reliability of our model.

# 8. Conclusion

This report aimed to provide a solution for Straits Times to streamline advertisements and conduct targeted marketing in-house especially in light of modern day privacy concerns. In this report, we addressed the business potential of a model to Straits Times, the issues that Straits Times faces with current advertising methods and the greater earning potential of using an in-house tailored marketing model. We also discussed data cleaning and explored interesting insights from the dataset. With this we analysed 5 different models using a balanced and unbalanced trainset before suggesting that the random forest model is best suited for the straits times to fulfil its objective, due to its relatively equal balance of FPR and FNR with a relatively low overall misclassification error. Random forest is also ideal as it has a greater explainability. Lastly, we suggested the ways to measure the success of the model before concluding with future actions to be taken to ensure sustainability and scalability of the recommended model.

# 9. Appendix

**Appendix A: Introduction**

****

*Figure 1.2: Image showing 3 different ads a reader of Straits Times will encounter*

## 

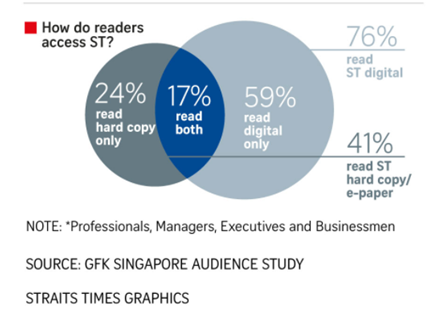
**Issues with Google Adsense and GoogleAds**

Straits Times currently uses Google Adsense to do advertising. Google Adsense uses cookies for advertising, including serving and rendering ads, personalising ads, limiting the number of times an ad is shown to a user, muting ads you have chosen to stop seeing, and measuring the effectiveness of ads. Without this information, Google Adsense may be unable to do accurate targeting, thereby reducing its effectiveness.

These privacy issues faced by GoogleAds can be circumvented through Singapore Press Holdings’ (SPH) authorised collection of and use of its subscribers’ personal data for advertising and marketing purposes. (*The SPH Media Privacy Policy states that subscribers must agree and consent to SPH Media Limited, its related corporations and affiliates, and their respective successors-in-title as well as our respective representatives collecting, using, disclosing and sharing amongst themselves your Personal Data, and disclosing such Personal Data to SPH Media’s authorised service providers and relevant third parties.*) Hence, SPH can adopt and implement its own digital marketing predictive analytics model for advertisements on its Straits Times website using their subscribers’ data, continuing to do accurate and targeted advertising.

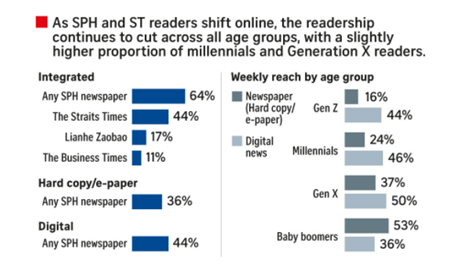
Google Adsense is also relatively expensive, with the average cost per click being USD0.63 for display advertisements. The average cost per click differs across different industries. ([Appendix A](#srnaiv5xkvhu)) A personalised model is therefore more price competitive and may potentially reduce the strain on the advertising budget.

**Appendix B: Case Justification**



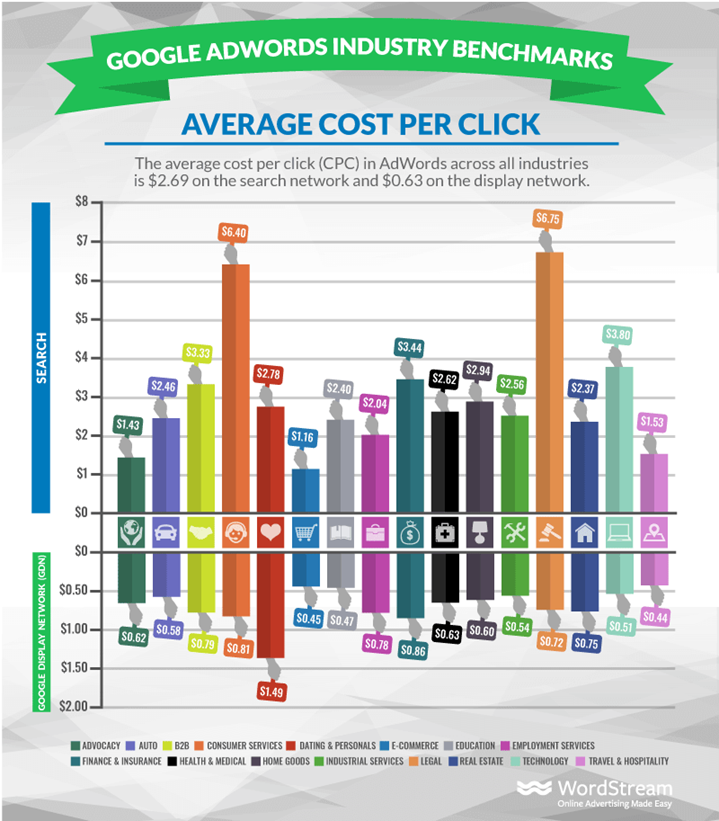
Source: The Straits Times

*Figure 2.1.1: Summary of readers' access to Straits Times*



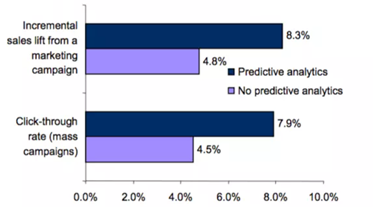
Source: The Straits Times

*Figure 2.1.2: Statistics of the demographic of Straits Times readers*



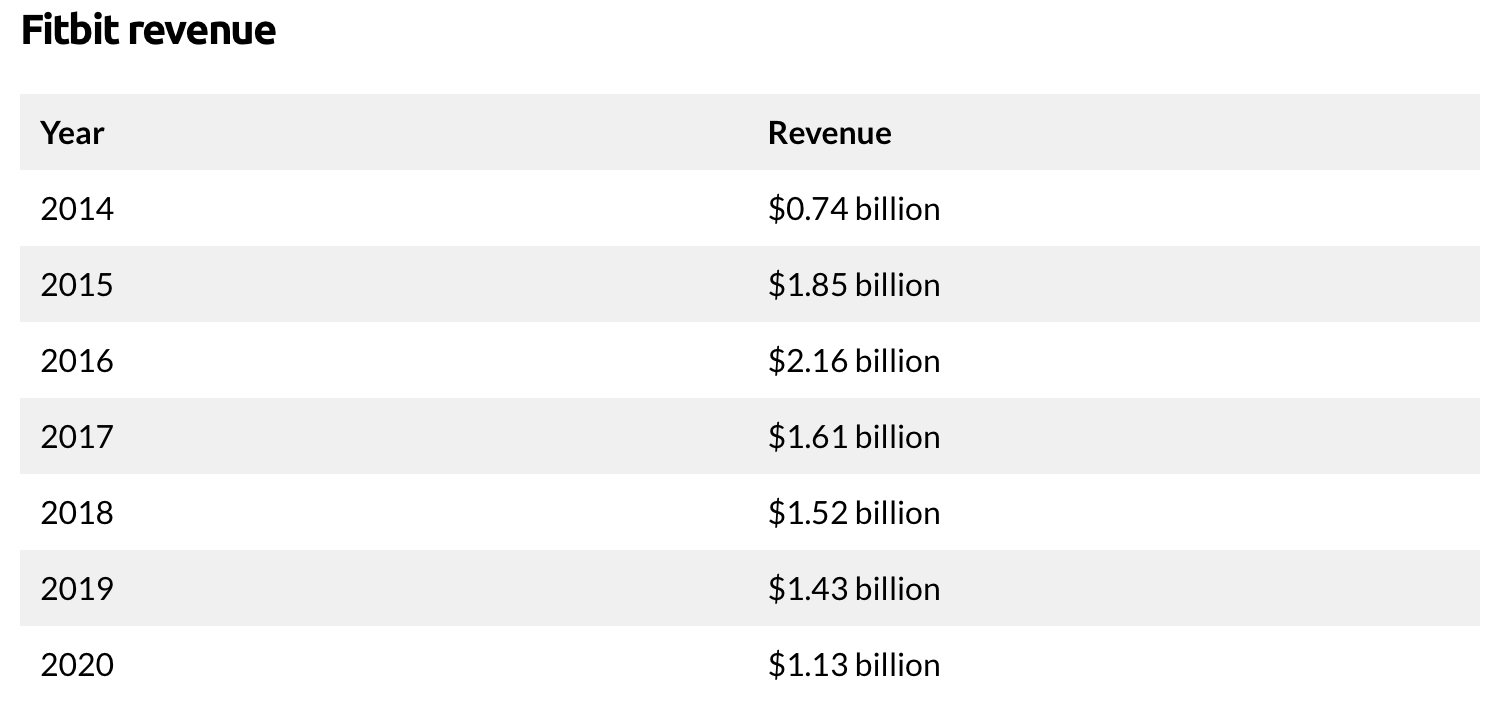
Source: Wordstream

*Figure 2.2.1: Statistics of the average cost per click of Google ads by industry*

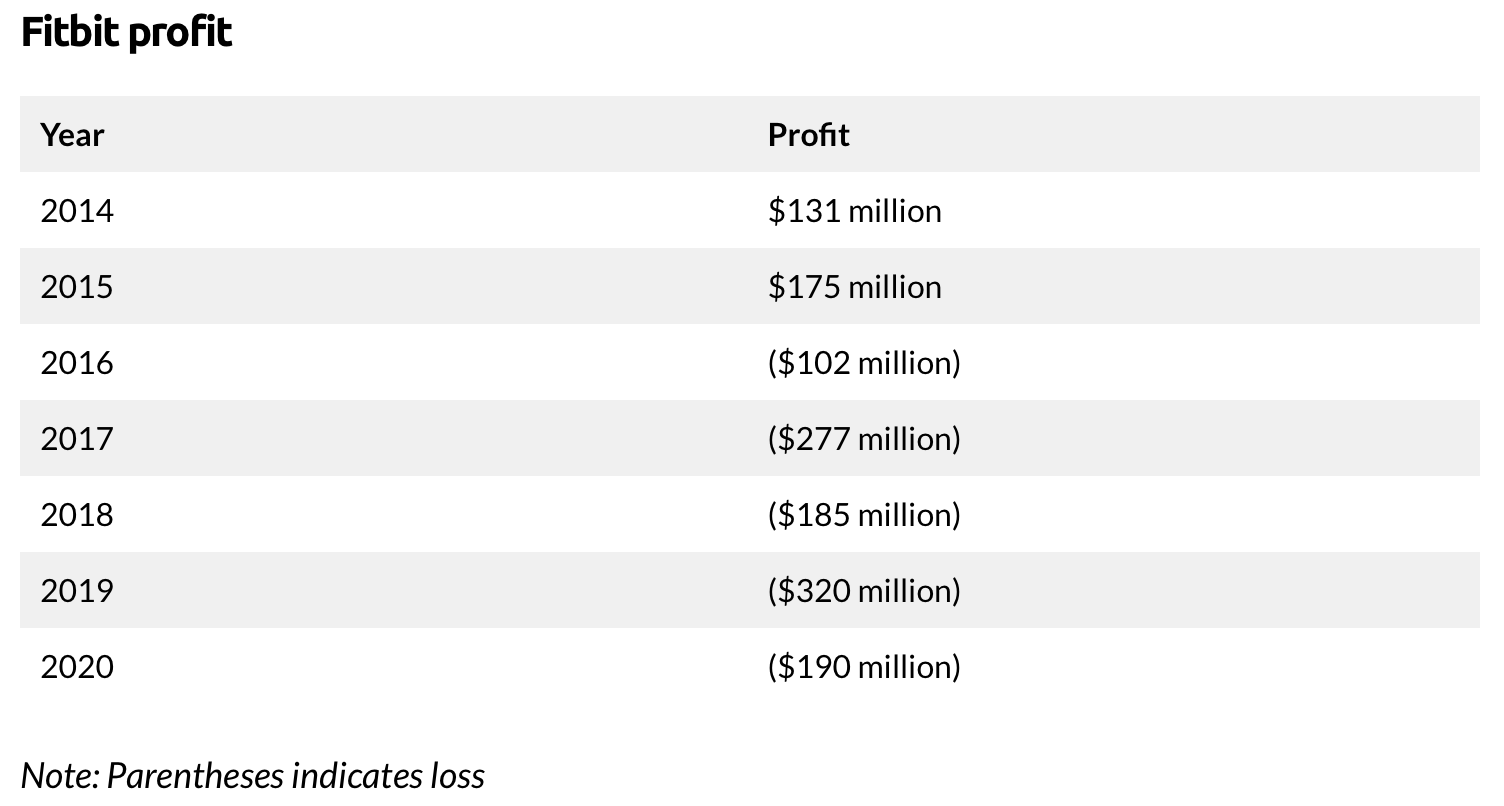


Source: Aberdeen Strategy and Research

*Figure 2.3.1: The impact of predictive analysis on sales and click-through rate*



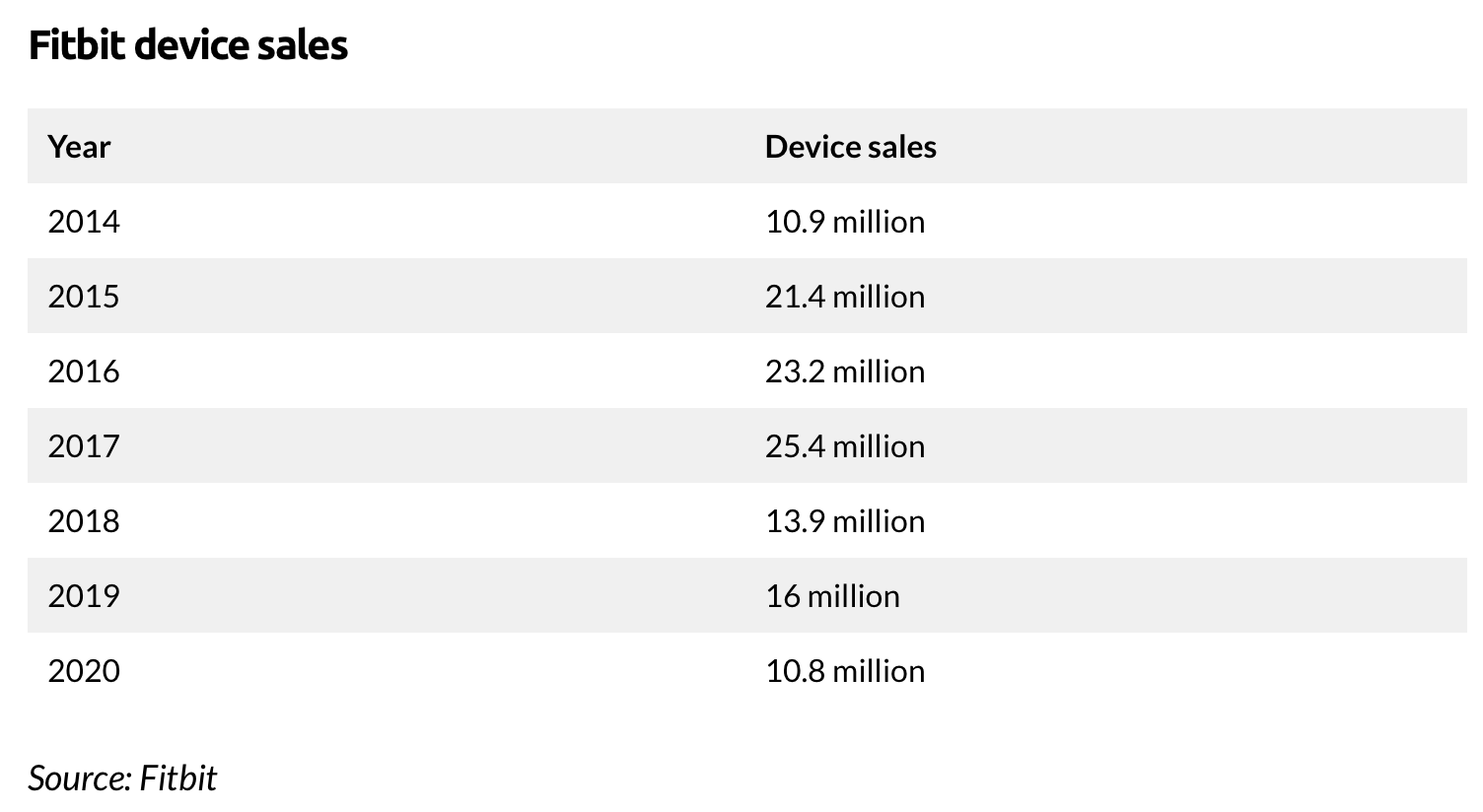
*Figure 2.4.1: Information on Fitbit’s Revenue over the years*



*Figure 2.4.2: Information on Fitbit’s profits over the years*

**

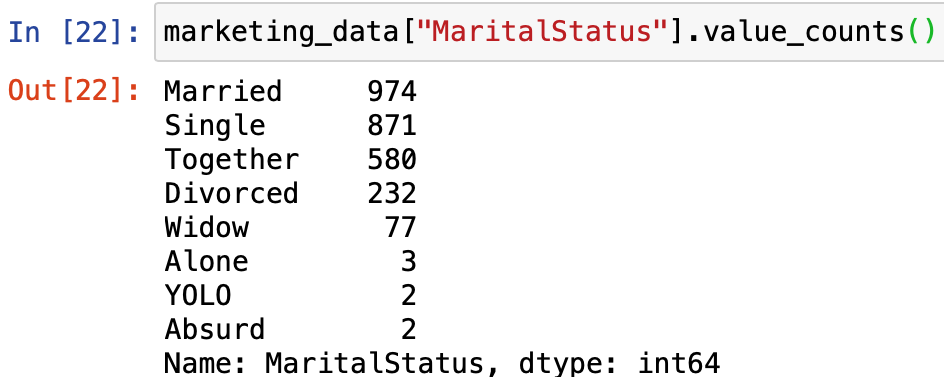
*Figure 2.4.3: Information on Fitbit’s users over the years*

**

*Figure 2.4.4: Information on Fitbit’s device sales over the years*

## 

**Appendix C: Data & Data Cleaning**



*Figure 3.1: Image showing some weird classification for Marital Status*



*Figure 3.2: Plot of inconsistencies in continuous values*

**Appendix D: Data Exploration**

Additional Data Visualisation

|  |  |
| --- | --- |
| **Diagram** | **Findings** |
|  | The proportion of people who clicked on the ads appears to be similar for both genders. The difference seems negligible so the relationship appears to be weak. |
|  | The proportion of people who clicked on the ads appears to be similar for all the different subscription payment methods. The difference seems negligible so the relationship appears to be weak |
|  | The mean income is similar for people who clicked on the ad and those who did not. However, the upper quartile is higher and the lower quartile is lower for those who clicked. The relationship appears to be weak. |
|  | The mean number of world events articles read is similar for the users who clicked on the ad and those who did not. However, the upper quartile is higher for those who clicked. The relationship appears to be weak. |

# 

Data Dictionary

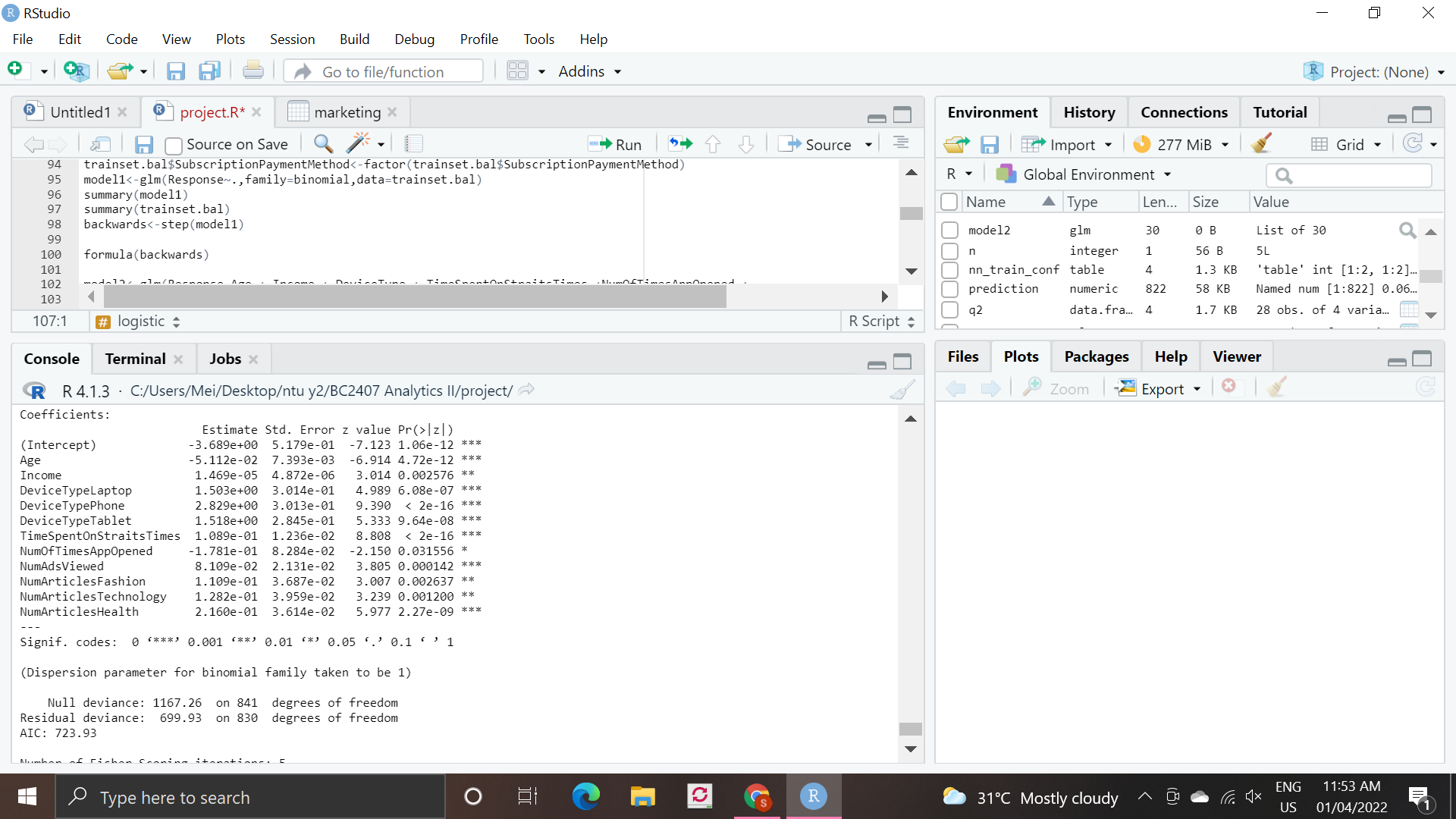
|  |  |
| --- | --- |
| **Variable name** | **Variable definition** |
| Age | The age of the reader |
| Gender | The gender of the reader.  0-Female  1-Male |
| ChildrenHome | Number of children in the household |
| Education | The highest academic qualification obtained by the reader.   * Basic refers to an individual who has obtained a high school education * Graduate refers to an individual who has obtained a basic college degree * Master refers to an individual who has obtained a masters degree * PhD refers to an individual who has obtained a Doctor of philosophy (PhD) * 2ndCycle refers to an individual who is on a second level of university studies |
| Marital Status | The reader’s marital status; married, single, divorced |
| Income | Customer’s yearly household income |
| DeviceType | Type of device used to access Straits Times app/ Straits Times digital; Phone, Laptop, Tablet, Computer |
| SubscriptionPaymentMethod | The mode of payment for ST monthly subscription  0- DBS GIRO  1- credit card  2- others |
| TimeSpentOnStraitsTimes | Duration spent on ST per day |
| NumOfTimesAppOpened | Number of times the user opens the ST app per day |
| NumAdsViewed | Number of advertisements shown to an individual user per day |
| NumArticlesFashion | Number of articles related to fashion read by an individual per day |
| NumArticlesTechnology | Number of articles related to technology read by an individual per day |
| NumArticlesHealth | Number of articles related to health read by an individual per day |
| NumArticlesSports | Number of articles related to sports read by an individual per day |
| NumArticlesWorldEvents | Number of articles related to world events read by an individual per day |
| Response | An individuals response to an advertisement  0- individual did not click on the ad  1- individual clicked on the ad |

## 

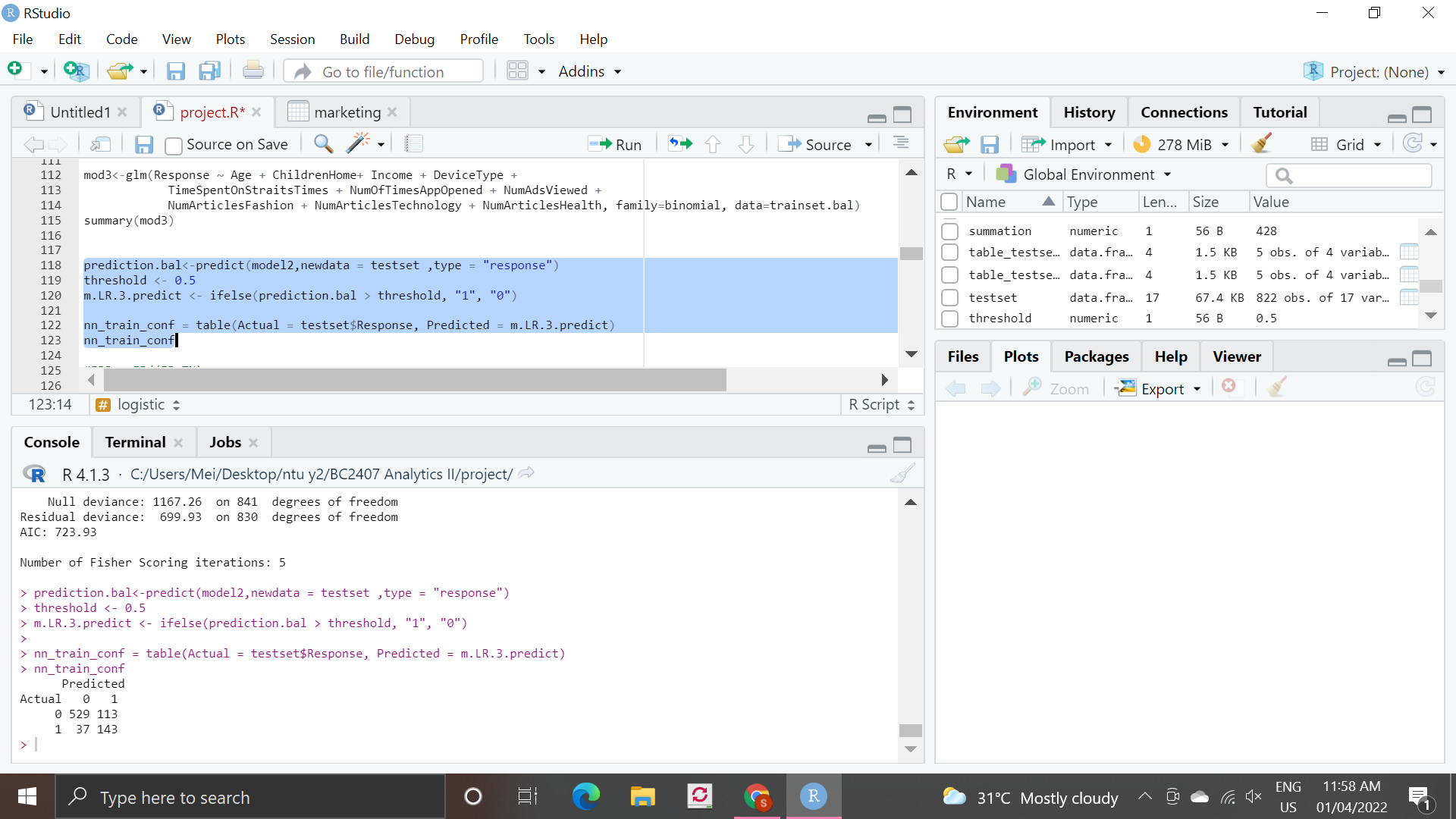
# 

**Appendix E: Analytics solution**

Logistic regression with backward elimination

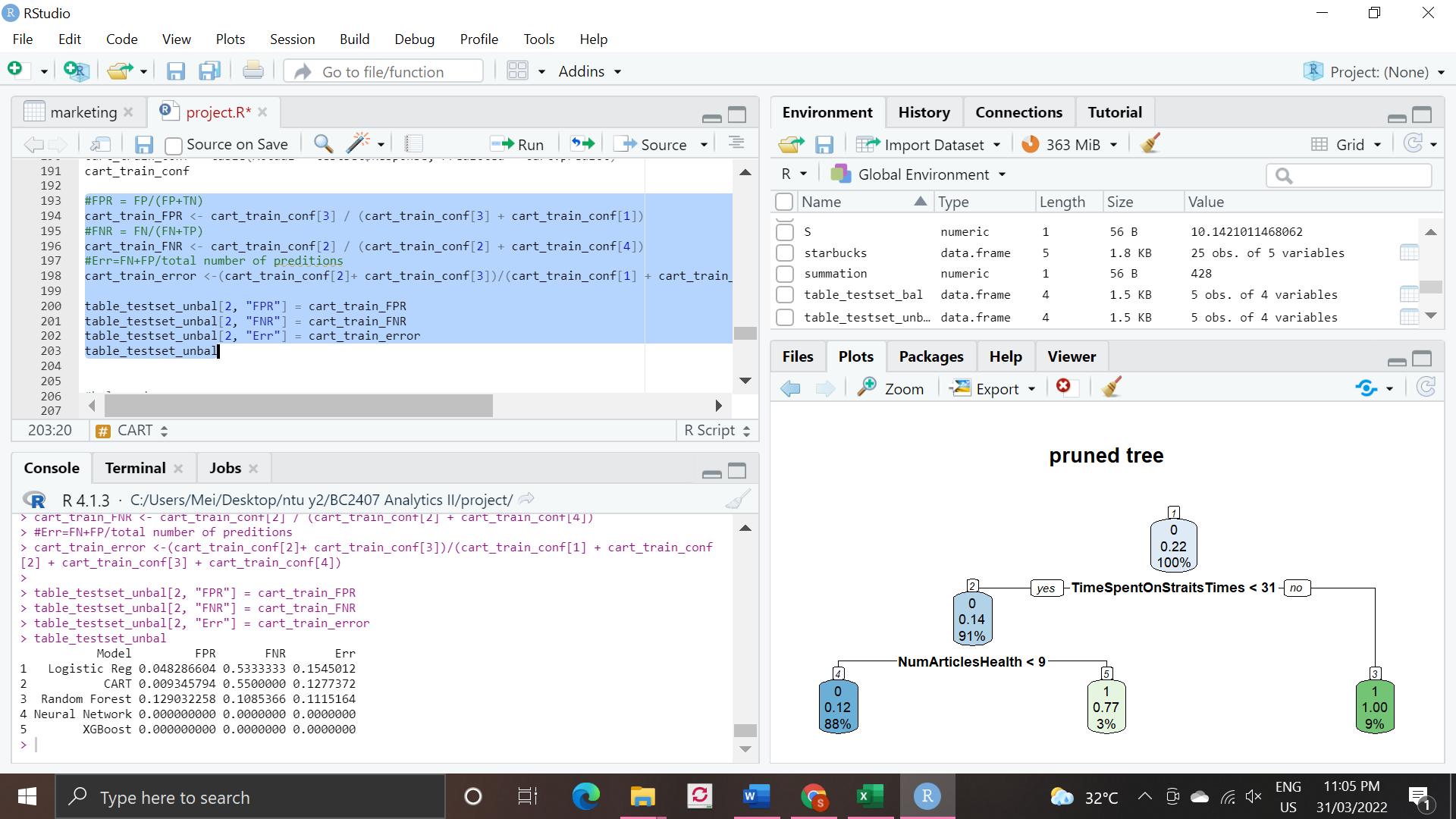


*Figure 5.3.3: Results obtained from logistic regression with backward elimination on the balanced trainset.*

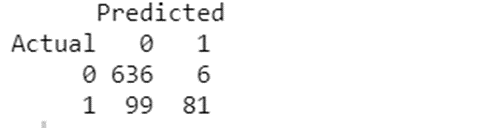


*Figure 5.3.4: Confusion matrix from logistic regression analysis on the balanced trainset.*

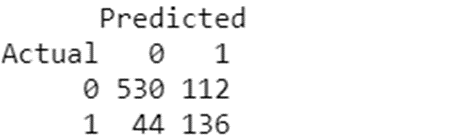
CART



*Figure 5.4.1: Decision tree from training on unbalanced dataset*

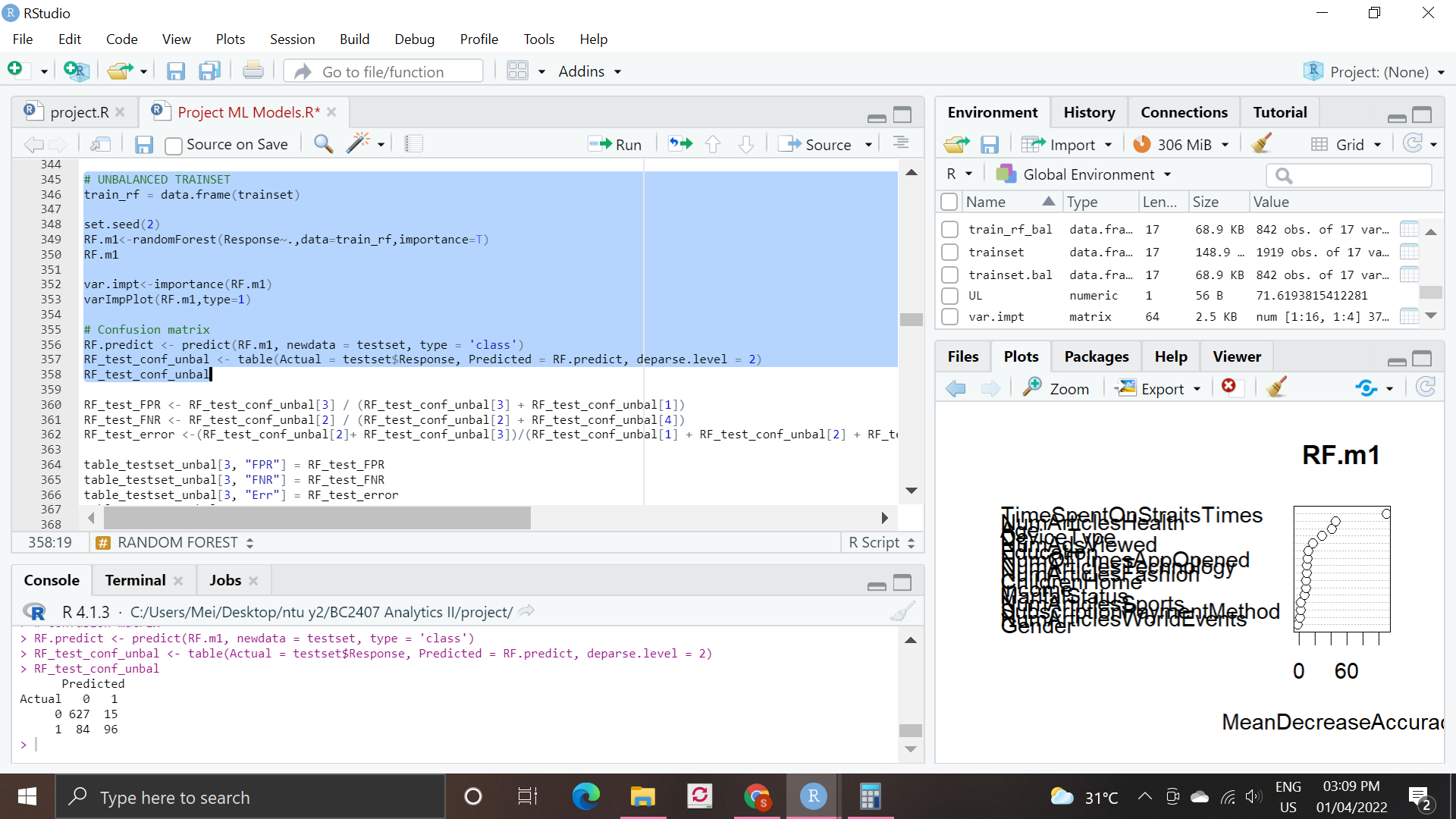


*Figure 5.4.2: Confusion matrix from training on unbalanced dataset*

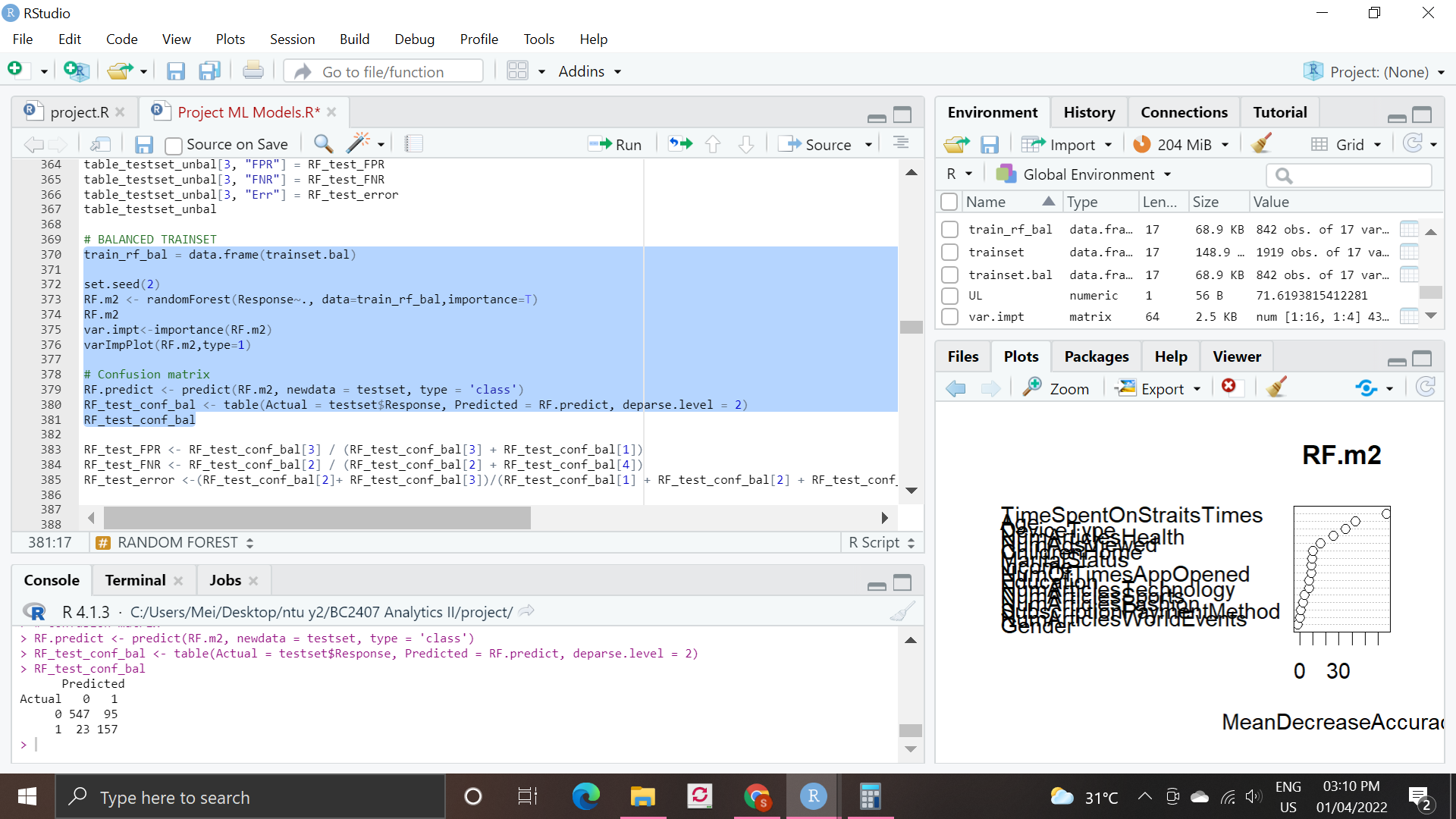


*Figure 5.4.4: Confusion matrix from training model on balanced trainset*

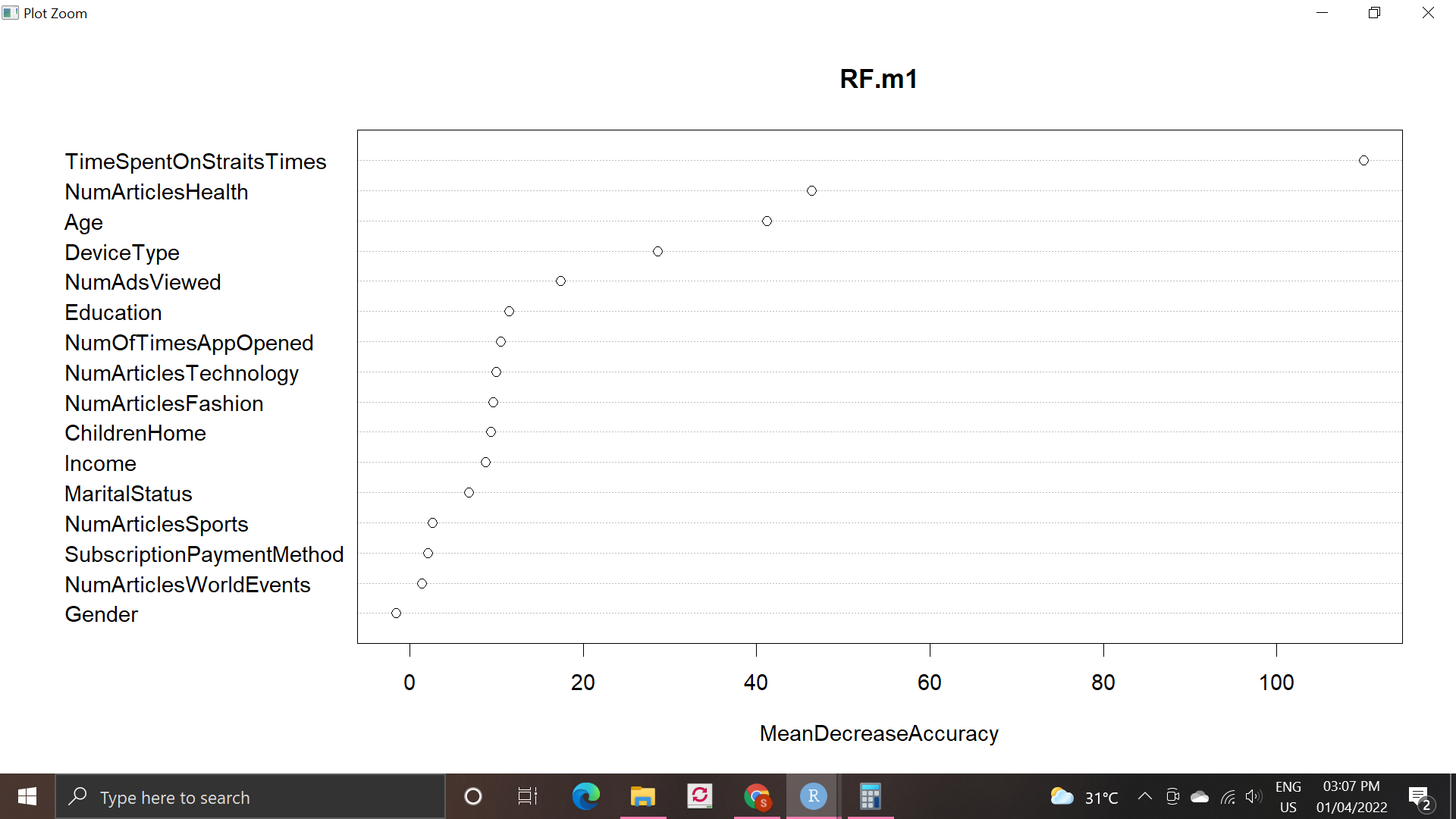
Random Forest



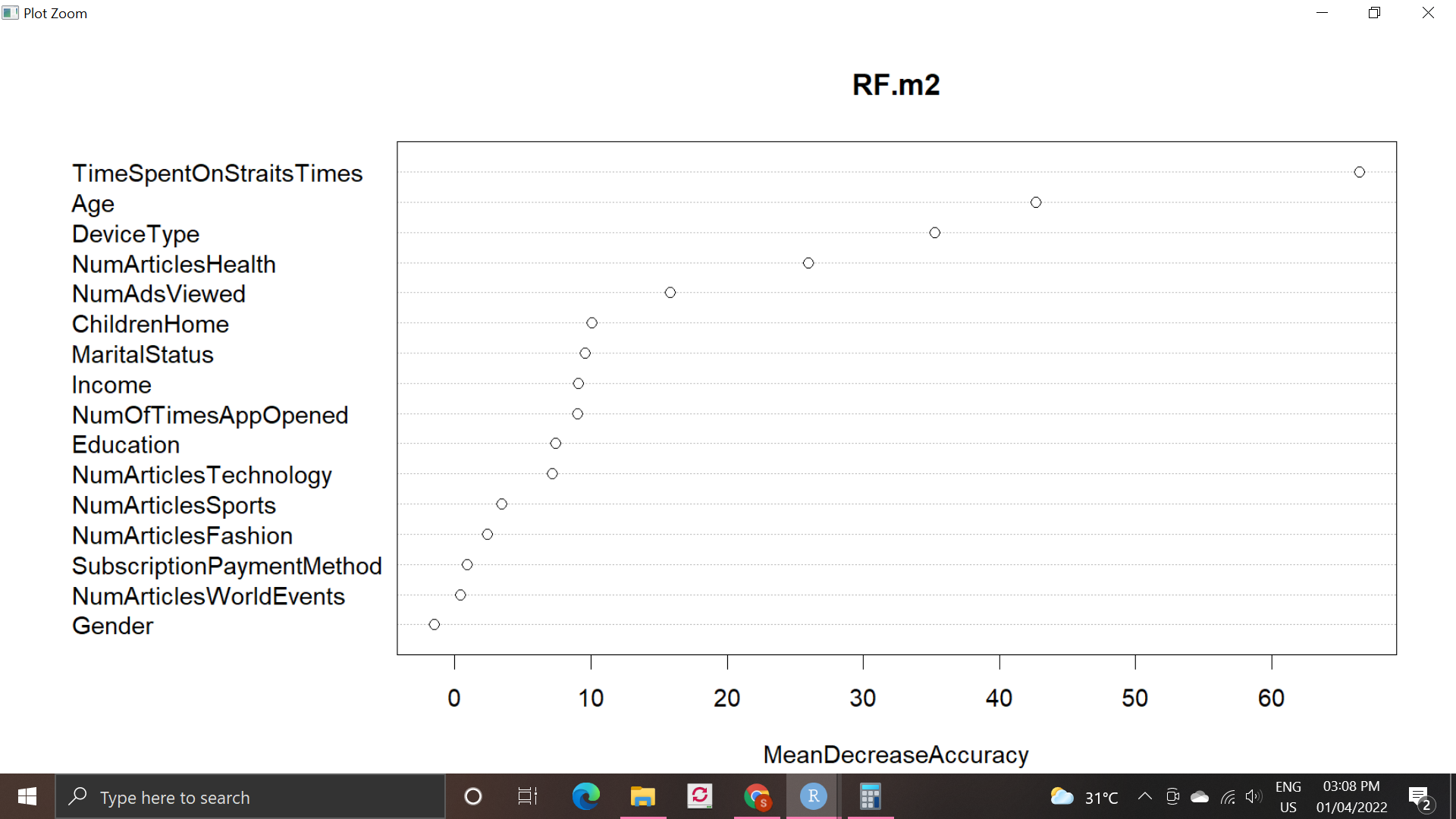
*Figure 5.5.1: Confusion matrix of unbalanced trainset*



*Figure 5.5.2: Confusion matrix of balanced trainset*



*Figure 5.5.3: Variable importance of unbalanced trainset*

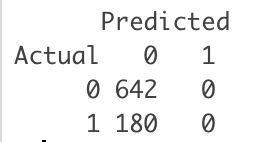


*Figure 5.5.4: Variable importance of balanced trainset*

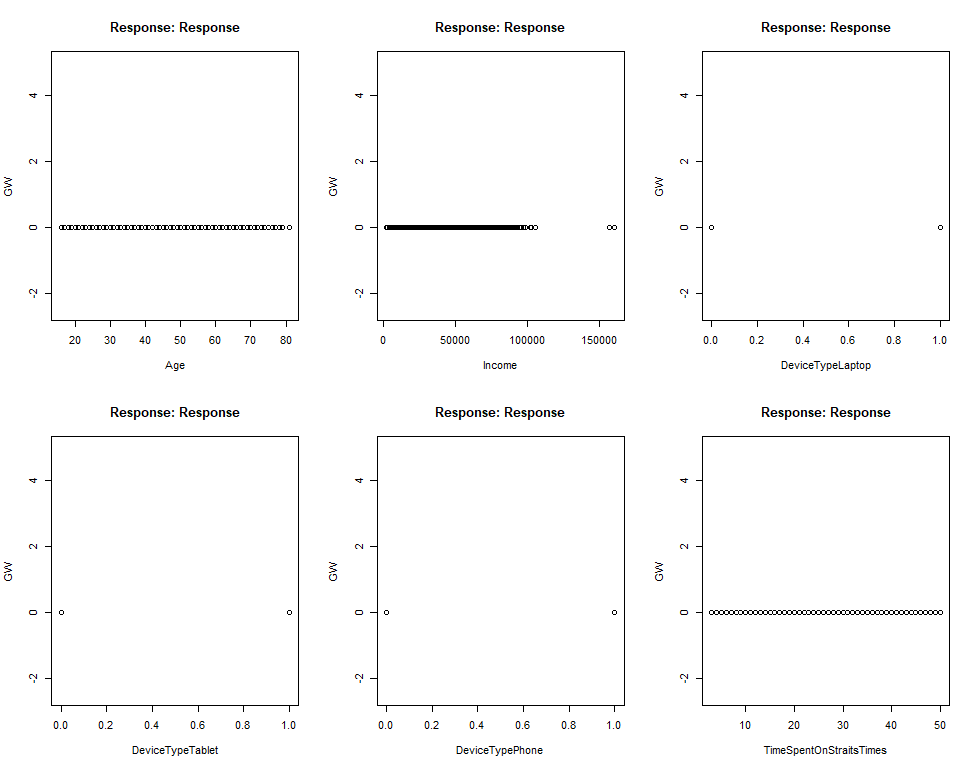
Neural Network

# 

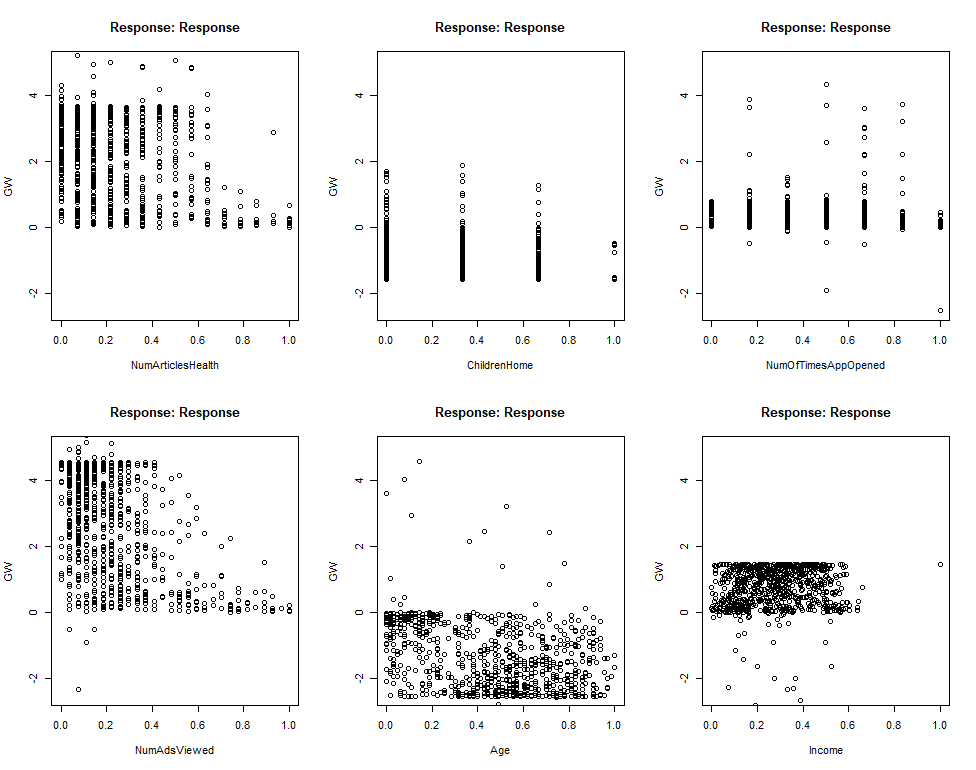
*Figure 5.6.1: Code to convert categorical into dummy variables using model.matrix()*



*Fig 5.6.2: Confusion Matrix for NN trained on unbalanced dataset*

**

*Figure 5.6.3: NN generalized weights plot against response (Without Normalization)*

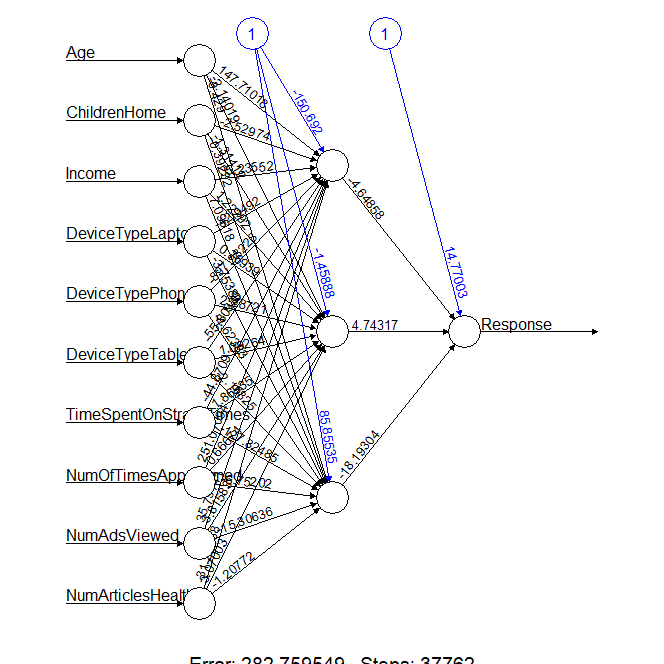
****

*Figure 5.6.6: NN generalized weights plot against response (With Normalization)*

# 

# 

# 

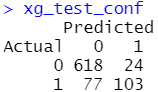


*Figure 5.6.7: NN diagram for model trained on balanced dataset with input variables from RF*

XGBoost

# 

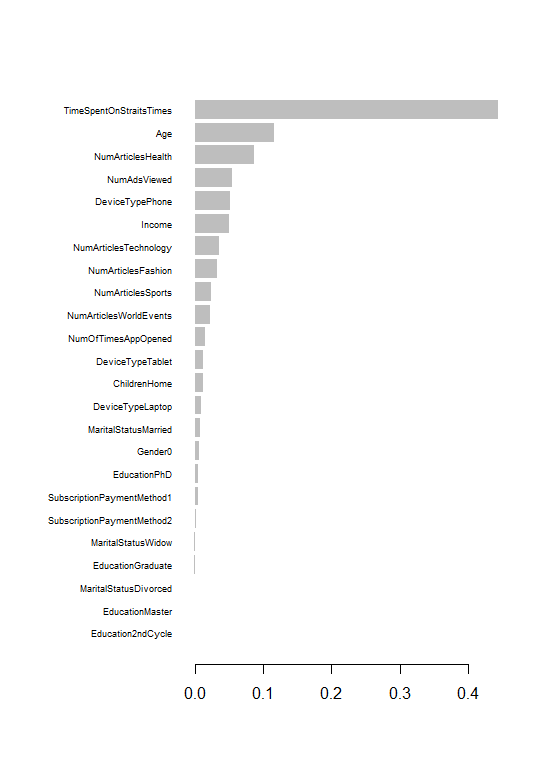
*Figure 5.7.1: NN diagram for model trained on balanced dataset with input variables from RF*



*Figure 5.7.3: Confusion matrix for XGBoost trained on unbalanced dataset (124th iteration)*

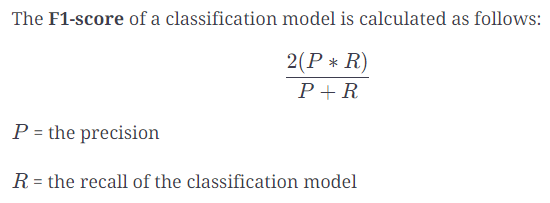


*Figure 5.7.4: Confusion matrix for XGBoost trained on balanced dataset (96th iteration)*

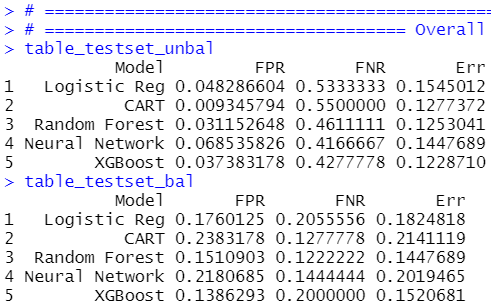


*Figure 5.7.5: Variable Importance plot for balanced dataset*

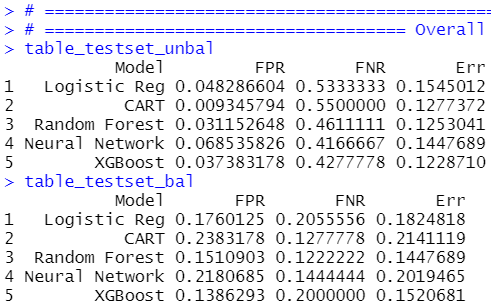
Model Recommendation



*Figure 5.8.1: Formula to calculate F1-Score*



*Figure 5.8.2: Table showing the error rates between different ML models (Unbalanced Dataset)*



*Figure 5.8.3: Table showing the error rates between different ML models (Balanced Dataset)*

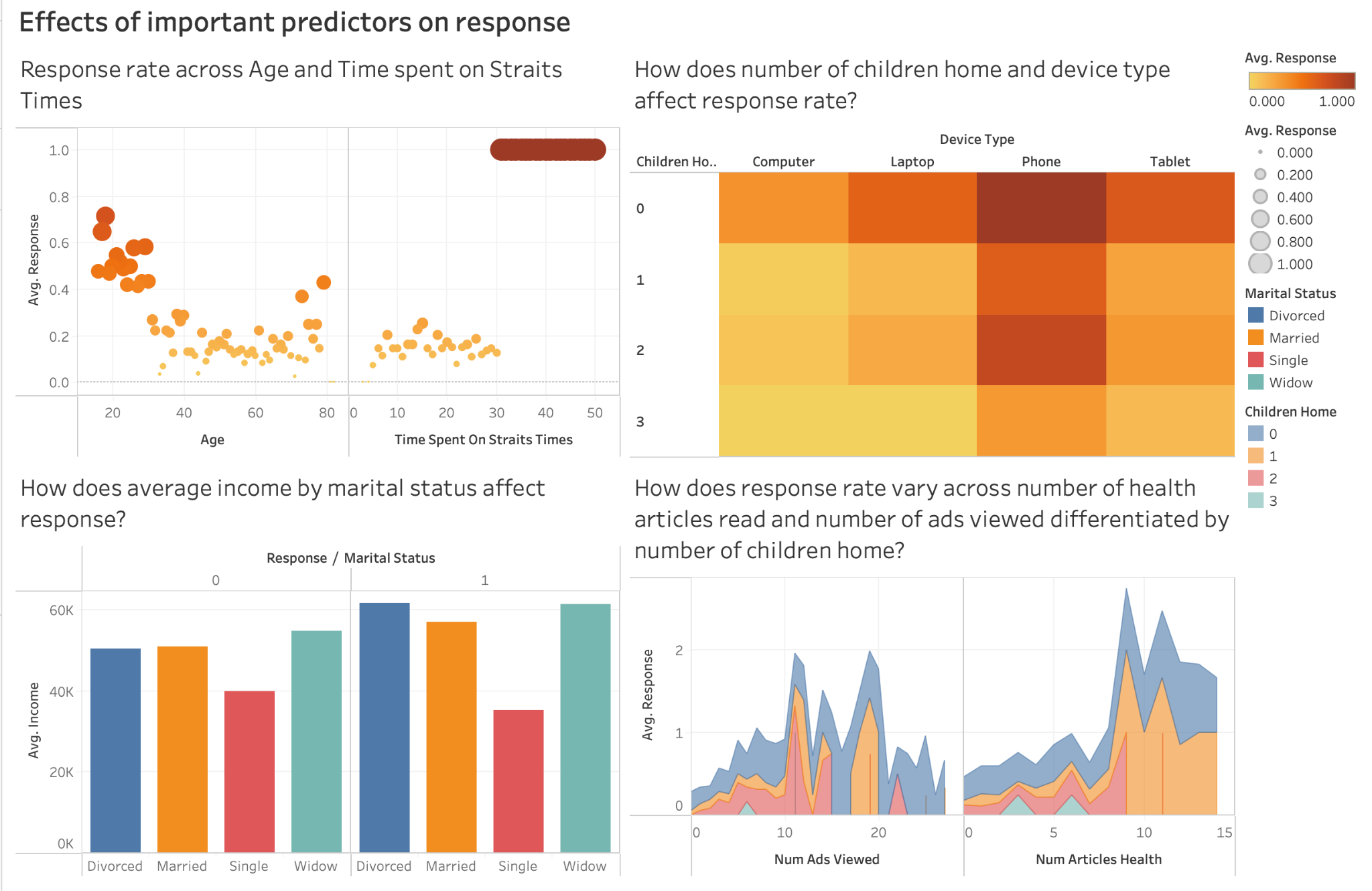
**Appendix F: Implementation Strategy**

## 

*Figure 6.1: Piechart showing the distribution of the most important variables*



*Figure 6.3: Dashboard exploring relationship between predictors*



*Figure 6.4: Dashboard exploring relationship between predictors*

# 10. References

1. Olsson, J. (2022, February 3). Google is Removing Third-Party Cookies in 2023. What’s Next? Match2One. <https://www.match2one.com/blog/how-removal-of-third-party-cookies-affects-digital-marketers/>
2. Internet Safety: Understanding Browser Tracking. (n.d.). GCFGlobal.Org. <https://edu.gcfglobal.org/en/internetsafety/understanding-browser-tracking/1/>
3. Bump, P. (2021, October 6). The Death of the Third-Party Cookie: What Marketers Need to Know About Google’s 2022 Phase-Out. Pamela Bump. <https://blog.hubspot.com/marketing/third-party-cookie-phase-out>
4. Olsson, J. (2022b, February 14). The Top 6 Digital Marketing Trends 2022. Match2One. <https://www.match2one.com/blog/6-programmatic-advertising-trends/>
5. Auto, H. (2021, September 20). *The Straits Times remains best-read title, with clear shift to digital, across all age groups: Survey*. The Straits Times. <https://www.straitstimes.com/singapore/community/the-straits-times-remains-best-read-title-with-clear-shift-to-digital-across-all>
6. Irvine, M. (2022, February 25). *Google Ads Benchmarks for YOUR Industry [Updated!]*. WordStream. <https://www.wordstream.com/blog/ws/2016/02/29/google-adwords-industry-benchmarks>
7. Chandrasekar, C. (2018, April 27). *What Is Predictive Advertising & Why Do You Need It?* Search Engine Journal. <https://www.searchenginejournal.com/predictive-advertising-need/248427/>
8. S., Meushar, S., K., & S. (2022, January 5). How to Make Money with Google AdSense. CodeFuel. <https://www.codefuel.com/blog/how-to-make-money-with-google-adsense/#:%7E:text=Google%20pays%20you%20per%20every,%243%20per%20click%20for%20publishers>.
9. Johnston, M. (2021, July 25). How Much Can Publishers Make from Taboola and Outbrain? MonetizePros. [https://monetizepros.com/display-advertising/which-sponsored-content-network-is-best-taboola-outbrain/#:%7E:text=In%20other%20 words%2C%20 most%20 publishers,meaningful%20revenue%20from%20these%20widgets](https://monetizepros.com/display-advertising/which-sponsored-content-network-is-best-taboola-outbrain/#:%7E:text=In%20other%20words%2C%20most%20publishers,meaningful%20revenue%20from%20these%20widgets).
10. Cooper, A. (2021, November 17). *Google cookie cull: how we can all approach data differently*. The Drum. <https://www.thedrum.com/news/2021/11/17/google-cookie-cull-how-we-can-all-approach-data-differently>
11. Hern, A. (2020, July 1). *Ministers urged to limit Facebook and Google’s power over ad market*. The Guardian. <https://www.theguardian.com/media/2020/jul/01/ministers-urged-to-limit-facebook-and-google-power-over-uk-ad-market>
12. Hern, A. (2021, January 25). Google announces plan to tackle privacy issues in online advertising. *The Guardian*. <https://www.theguardian.com/technology/2021/jan/25/google-announces-plan-to-tackle-privacy-issues-in-online-advertising>
13. Auto, H. (2021, July 19). *Operating revenue of SPH's media business falls for first 9 months of FY21*. The Straits Times. <https://www.straitstimes.com/business/companies-markets/operating-revenue-of-sphs-media-business-falls-for-first-9-months-of-fy21>
14. Gogoi, K. (2021, December 8). *9 Modern Day Digital Marketing Challenges fro E-commerce And How To Overcome Them*. Wigzo. <https://www.wigzo.com/blog/digital-marketing-challenges/>
15. S. (2021, January 15). *Predictive analytics in marketing: everything you need to know*. Supermetrics. <https://supermetrics.com/blog/predictive-analytics-in-marketing>
16. Convergence, I. T. (2021, October 13). *A Complete Guide to Predictive Analytics*. IT Convergence. <https://www.itconvergence.com/blog/a-complete-guide-to-predictive-analytics/>
17. Morrison, S. (2021, March 3). *Google’s third-party cookie ban, explained*. Vox. <https://www.vox.com/recode/2021/3/3/22311460/google-cookie-ban-search-ads-tracking>
18. Sachdev, H. S. (2021, April 20). Choosing number of Hidden Layers and number of hidden neurons in Neural Networks. LinkedIn. <https://www.linkedin.com/pulse/choosing-number-hidden-layers-neurons-neural-networks-sachdev/>
19. *Fitbit Revenue and Usage Statistics (2022)*. (2022, January 11). Business of Apps. <https://www.businessofapps.com/data/fitbit-statistics/>
20. Papadopoulos, S. (2009). Key Success Factors in Internet Advertising. *Key Success Factors in Internet Advertising*. <https://www.diva-portal.org/smash/get/diva2:830978/FULLTEXT01.pdf>
21. Wikipedia contributors. (2022, March 27). *Fitbit*. Wikipedia. <https://en.wikipedia.org/wiki/Fitbit>
22. <https://www.businesstimes.com.sg/opinion/a-thriving-local-news-media-in-the-digital-age>
23. <https://corporate.sph.com.sg/legal/sph_privacy.html>